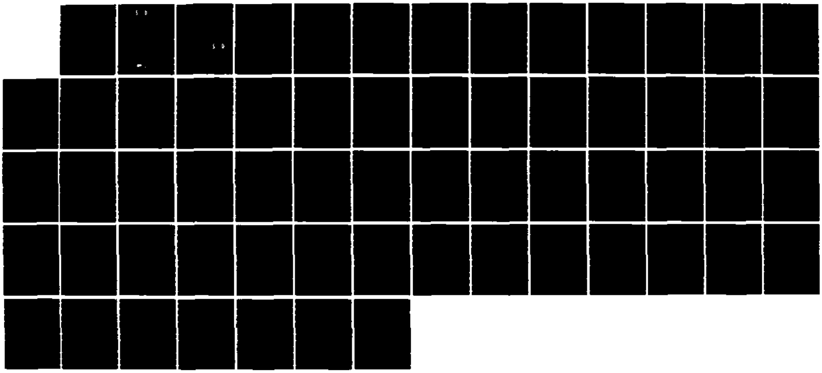
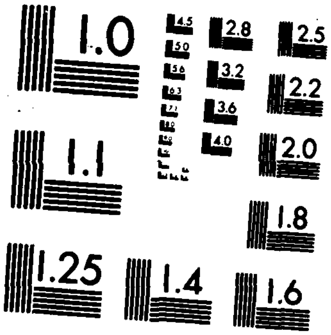
AD-A173 201 PROBABILISTIC ANALYSIS OF A RELAXATION FOR THE K-MEDIAN 1/1
PROBLEM REVISION(U) CARNEGIE-MELLON UNIV PITTSBURGH PA
MANAGEMENT SCIENCES RESEAR.. S AHN ET AL. JUN 86
UNCLASSIFIED MSRR-527 N00014-85-K-0198 F/G 12/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

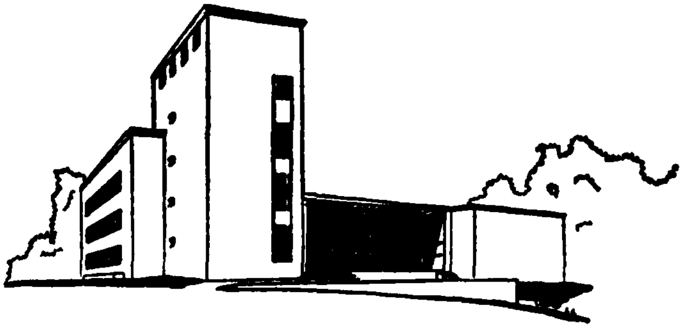
1

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

GRADUATE SCHOOL OF INDUSTRIAL ADMINISTRATION

WILLIAM LARIMER MELLON, FOUNDER



1

Management Science Research Report No. MSRR 527

PROBABILISTIC ANALYSIS OF A RELAXATION FOR THE k-MEDIAN PROBLEM

by

Sang Ahn[1]

Colin Cooper[2]

Gerard Cornuejols[3]

Alan Frieze[4]

May 1985
(Revised June 1986)

[1] Seoul National University, Seoul, S. Korea.

[2] Polytechnic of North London, London, England.

[3] Carnegie Mellon University, Pittsburgh, USA.

[4] Queen Mary College, London, England.

This report was prepared in part as part of the activities of the Management Sciences Research Group, Carnegie Mellon University, under Contract No. N00014-85-K-0198 NR 047-048 with the Office of Naval Research and in part by an NSF grant ECS8503192. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

Management Sciences Research Group
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

This paper provides a probabilistic analysis of the so-called "strong" linear programming relaxation of the k-median problem. The analysis is performed under four classical models in location theory, the Euclidean, network, tree and uniform cost models. For example, it is ~~we~~ shown that, for the Euclidean model and $\log n \leq k \leq n/(\log n)^2$ (≤ or =, Squared), the value of the relaxation is almost surely within .3 percent of the optimum k-median value. A similar analysis is performed for the other models. It is ~~We~~ also shown that, under various assumptions, branch and bound algorithms that use this relaxation as a bound must almost surely expand a non-polynomial number of nodes to solve the k-median problem to optimality. Finally, ~~we report~~ extensive computational experiments are reported. As predicted by the probabilistic analysis, the relaxation was not as tight for the problem instances drawn from the uniform cost model as for the other models.

1. Introduction

The k-median problem has been widely studied both from the theoretical point of view and for its applications. An interesting theoretical development was the successful probabilistic analysis of several heuristics for this problem (e.g. Fisher and Hochbaum [8] and Papadimitriou [22]). On the other hand, the literature on the k-median problem abounds in exact algorithms. Most are based on the solution of a certain relaxation to be defined later. The computational experience reported in the literature seems to indicate that this particular relaxation yields impressively tight bounds compared to what can usually be expected in integer programming. In this paper we analyze to what extent this relaxation is tight. We perform our analysis under various probabilistic assumptions and identify conditions under which the relaxation can be expected to be tight and others under which it can be expected to give a poor bound. For example, for a classical Euclidean model in the plane, we show that the relaxation can be expected to provide a bound within one third of one percent of the optimum value of the k-median problem. In addition to the probabilistic analysis, we also report extensive computational experiments, based on the solution of thousands of medium-size problems. Some of the results predicted for very large problems by our probabilistic analysis can already be observed on these test problems.

Consider a set $X=\{X_1,\dots,X_n\}$ of n points, a positive integer $k \leq n$ and let $d_{ij} \geq 0$ be the distance between X_i and X_j for each $1 \leq i \leq n$ and $1 \leq j \leq n$. (Unless otherwise specified, it is assumed that $d_{ii} = 0$, $d_{ij} = d_{ji}$ and $d_{ij} \leq d_{ik} + d_{kj}$, for all i,j,k). The _k-median problem_ consists of finding a set $S \subseteq X$, $|S| = k$, that minimizes $\sum_{i=1}^{n} \min_{j\epsilon S} d_{ij}$. (Here $|S|$ denotes the cardinality of the set S.) The k-median problem has the following integer programming formulation.

$$(1) \qquad z_{IP} = \min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij}$$

$$(2) \qquad \sum_{j=1}^{n} y_{ij} = 1 \quad \text{for } i=1,\dots,n$$

$$(3) \qquad \sum_{j=1}^{n} x_j = k$$

$$(4) \qquad 0 \leq y_{ij} \leq x_j \leq 1 \text{ for } i,j = 1,\dots,n$$

$$(5) \qquad x_j \varepsilon \{0,1\} \text{ for } j = 1,\dots,n.$$

In this formulation $x_j = 1$ if $X_j \varepsilon S$, 0 otherwise and, for $1 \leq i \leq n$, we can set $y_{ij} = 1$ for an index j that achieves $\min_{j\varepsilon S} d_{ij}$.

The formulation (1)-(4) is called the linear programming (LP) relaxation of the k-median problem. In other words, the LP relaxation is obtained by ignoring the integrality conditions on x_j, $1 \leq j \leq n$. The optimum value z_{LP} of this relaxation clearly satisfies $z_{LP} \leq z_{IP}$. The bound z_{LP} has been used extensively in exact algorithms for the k-median problem. (E.g. Marsten [15], Garfinkel Neebe and Rao [10], ReVelle and Swain [23], Diehr[5], Shrage[24], Guignard and Spielberg[11], Narula, Ogbu and Samuelsson[20], Cornuejols, Fisher and Nemhauser [3], Erlenkotter [6], Galvão [9], Magnanti and Wong [14], Nemhauser and Wolsey [21], Mulvey and Crowder [19], Mavrides [16], Mirchandani, Oudjit and Wong [17], Christofides and Beasley [2], Beasley[1].)

Most of the computational experience has been reported on test problems with $n \leq 100$. For many of these test problems, $z_{IP} = Z_{LP}$. Recently, Beasley [1] solved forty larger problems (with $100 \leq n \leq 900$) and found a small but positive gap $z_{IP}-z_{LP}$ for many of them. The average of $\frac{z_{IP} - z_{LP}}{z_{IP}}$ over these problems was .0024.

In this paper we analyze the ratio $\frac{z_{IP}-z_{LP}}{z_{IP}}$ from a probabilistic point of view as n goes to infinity, under various assumptions on the probability distribution of problem instances. We do not address the worst-case analysis of this ratio except to note that this question was solved by Cornuejols, Fisher and Nemhauser [3] when $d_{ij} \le 0$. The analysis of [3] does not carry over when the d_{ij}'s are nonnegative and satisfy the distance axioms. In fact, this worst-case analysis is an interesting open question. It would also be interesting to know the worst-case value of $\frac{z_{IP}-z_{LP}}{z_{IP}}$ when the d_{ij}'s are further restricted to represent Euclidean distances. Once again, these questions are not addressed here as we focus on a probabilistic approach.

We will often write statements like $X_n \le u_n$ almost surely (a.s.) for a sequence of random variables (X_n) and real sequence (u_n). This is a well-defined terminology of probability theory and details can be found in Stout [25] for example. We will invariably prove that

$$\sum_{n=1}^{\infty} \Pr(X_n > u_n) < \infty$$

which implies the above statement. Non-probabilists will be satisfied that we show $\Pr(X_n > u_n) \to 0$ as $n \to \infty$. If $X_n \le u_n(1+o(1))$ a.s. and $X_n \ge u_n(1-o(1))$ a.s. then we write $X_n \sim u_n$ a.s.

First we study the k-median problem in the plane. When the points $X_1,\dots,X_n$ are uniformly distributed in a unit square and d_{ij} is the Euclidean distance between X_i and X_j, $1 \le i,j \le n$, we show that $\frac{z_{IP}-z_{LP}}{z_{IP}} \sim .00284$ almost surely, for any k such that $\omega \le k \le \frac{n}{\omega \log n}$ where $\omega = \omega(n) \to \infty$. (In this paper we abbreviate $f(n) \to a$ as $n \to \infty$ by $f(n) \to a$.)

In a second model, the points $X_1,\dots,X_n$ are the nodes of a random graph $G_n(p)$ where p is the probability that an edge is in the graph, and d_{ij} is the number of edges on the shortest path from X_i to X_j. We assume $p \ge \frac{\omega \log n}{n}$ where $\omega = \omega(n) \to \infty$ (this guarantees that $G_n(p)$ is almost surely connected), and $kp^2 \ge \frac{\omega \log n}{n}$. We prove that $\frac{z_{IP} - z_{LP}}{z_{IP}} \le \frac{1}{e+1}$ almost surely, where e is the base of natural logarithms. More specifically, if $\log_b n \le k \le n$ where $b = \frac{1}{1-p}$, then $z_{IP} = z_{LP}$ almost surely. If $2 \le k \le \log_b n$, $kp \to \alpha$ where $0 \le \alpha \le \infty$ and $p \to \beta$ where $0 \le \beta < 1$, define $a \equiv e$ if $\beta = 0$ and $(1-\beta)^{1-\beta}$ if $\beta > 0$; then $\frac{z_{IP} - z_{LP}}{z_{IP}} \sim f(\alpha,\beta)$ almost surely where $f(\alpha,\beta) = \frac{1-(1-\alpha)^{+}a^{\alpha}}{1+a^{\alpha}}$. (The maximum of this function is $\frac{1}{e+1}$ attained when $\alpha=1$ and $\beta=0$. When $\alpha=0$ or ∞ the function takes the value 0).

We also analyze the k-median problem on random trees and on another model where it is assumed that the d_{ij}'s are independently and uniformly distributed on [0,1].

In section 6, we put our probabilistic results in perspective by presenting extensive computational experiments.

In section 7, we show how our results for the k-median problem relate to the simple plant location problem (SPLP). In the SPLP, the data comprise n points $X_1, \ldots, X_n$, distances d_{ij} for $1 \le i, j < n$, and fixed costs f_j associated with each point X_j, $1 \le j \le n$. The SPLP consists of finding a nonempty set $S \subseteq X$ that minimizes $\sum_{i=1}^{n} \min_{j \varepsilon S} d_{ij} + \sum_{j \varepsilon S} f_j$. (Note that, in this problem, $|S|$ is not restricted as in the k-median problem.) An integer programming formulation of SPLP is

$$z_{IP} = \min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij} + \sum_{j=1}^{n} f_j x_j$$

subject to (2), (4) and (5). The LP relaxation is obtained by relaxing the integrality conditions (5).

In the remainder of this section we state some useful results from the literature. Our proofs use the following lemma (see Hoeffding[12]).

Lemma 1. If $Y_1, \ldots, Y_n$ are independent random variables and $0 \le Y_i \le 1$ for

$i=1,\ldots,n$, then, for $0<\varepsilon<1$,

$$\Pr(\bar{Y} \geq (1+\varepsilon)\mu) \leq e^{-\varepsilon^2 n\mu/3} \quad \text{and}$$

$$\Pr(\bar{Y} \leq (1-\varepsilon)\mu) \leq e^{-\varepsilon^2 n\mu/2} ,$$

where $\bar{Y} = (\sum_{i=1}^{n} Y_i)/n$ and μ is the expected value of $\bar{Y}$.

Given a vector $x = (x_j : j=1,\ldots,n)$ such that $\sum_j x_j = k$ and $0 \leq x_j \leq 1$ for all j, define

$$\begin{aligned} z_{LP}(x) = \min \ & \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij} \\ & \sum_{j=1}^{n} y_{ij} = 1 \text{ for } i=1,\ldots,n \\ & 0 \leq y_{ij} \leq x_j \text{ for } i,j,=1,..,n. \end{aligned}$$

Note that $z_{LP} = \min z_{LP}(x)$

$$\begin{aligned} & \sum_j x_j = k \\ & 0 \leq x_j \leq 1 \text{ for } j=1,\ldots,n. \end{aligned}$$

The following lemma is well-known in the k-median literature and is easy to prove.

<u>Lemma 2</u> An optimal solution $y = (y_{ij}: i,j = 1,\ldots,n)$ of $z_{LP}(x)$ is obtained as follows. For each i, sort the values d_{ij}, $j=1,\ldots,n$, so that

$$d_{ij_1(i)} \leq d_{ij_2(i)} \leq \ldots \leq d_{ij_n(i)},$$

and let p be such that $\sum_{h=j_1(i)}^{j_{p-1}(i)} x_h \leq 1 \leq \sum_{h=j_1(i)}^{j_p(i)} x_h$.

Then

$$y_{ij} = \begin{cases} x_j & \text{for } j = j_1(i),\dots,j_{p-1}(i) \\ 1 - \sum_{h=j_1(i)}^{j_{p-1}(i)} x_h & \text{for } j = j_p(i) \\ 0 & \text{for } j = j_{p+1}(i),\dots,j_n(i) . \end{cases}$$

<u>Proof</u>. The program $z_{LP}(x)$ separates for each j into a linear program with upper bounded variables and a single constraint.

□

Let $d_i = \sum_{j=1}^{n} d_{ij}y_{ij}$ where the values of y_{ij} are those defined in Lemma 2. Note that $z_{LP} \le \sum_{i=1}^{n} d_i$ since this bound is derived from a primal feasible solution. This bound will be used repeatedly in our proofs where it is computed for the vector x defined by $x_j = k/n$ for j=1,...,n.

The dual of the LP relaxation is

$$(6) \qquad z_{LP} = \max \sum_{i=1}^{n} u_i - \sum_{j=1}^{n} v_j - kw$$

$$u_i - t_{ij} \le d_{ij} \text{ for all } i,j$$

$$\sum_{i=1}^{n} t_{ij} - v_j - w \le 0 \text{ for all } j$$

$$t_{ij}, v_j \ge 0 \text{ for all } i,j.$$

For any given vector $u = (u_i : i=1,\dots,n)$, define

$$\rho_j(u) = \sum_{i=1}^{n} (u_i - d_{ij})^+ \quad \text{for } j = 1,\ldots,n,$$

where a^+ denotes $\max(0,a)$. Let $z_D(u) = \sum_{i=1}^{n} u_i - k \max_{j=1,\ldots,n} \rho_j(u)$.

Lemma 3. $z_{LP} \geq z_D(u)$ for any vector u.

Proof: It can be checked that, for any given u, a feasible solution of (6) is obtained by setting $t_{ij} = (u_i - d_{ij})^+$, $v_j = 0$ and $w = \max_{j=1,\ldots,n} \rho_j(u)$.

□

2. The Euclidean model in the plane.

This section is concerned with the following Euclidean model: n points $X_1,\ldots,X_n$ are chosen independently and uniformly at random in the unit square $S_0 = [0,1]^2$. The distance matrix is given by $d_{ij} = ||X_i - X_j||$ for $1 \leq i,j \leq n$ where $||\bullet||$ denotes the Euclidean norm. We assume that

(7) $\quad k \to \infty$ and $n/(k\log n) \to \infty$.

The following theorem was proved by Papadimitriou [22].

Theorem 1 Under the above conditions,

$$z_{IP} \sim (.3771967\ldots)\, n/\sqrt{k} \text{ a.s.}$$

This result was obtained by comparing z_{IP} to the value z_C of finding k points in $X = \{X_1,\ldots,X_n\}$ that minimize the sum of the distances to a continuum of points in the unit square. Papadimitriou showed that, when (7) holds, $z_{IP} \sim z_C$ almost surely. Actually, he used a weaker notion of probabilistic convergence, but Zemel [26] showed that almost sure convergence holds as well. It should be pointed out, however, that the continuous problem yielding z_C is very different from the LP relaxation. In fact, for the LP relaxation, we prove

Theorem 2 Under the above conditions,

$$z_{LP} \sim \frac{2}{3\sqrt{\pi}} \; n/\sqrt{k} \quad \text{a.s.}$$

where $2/(3\sqrt{\pi}) = .3761264\ldots.$

Our method of proof consists of conjecturing a near-optimal solution to the LP relaxation and a near-optimal solution to its dual. Then we show that, almost surely, these lower and upper bounds on z_{LP} are the same, up to small order terms. The probabilistic arguments are based on the estimates of the tails of the binomial distribution given in Lemma 1.

The proof of Theorem 2 will actually provide a constructive way of obtaining an upper bound $z_{LP}(x)$ and a lower bound $z_D(u)$ on the optimum value of the LP relaxation of the k-median problem.

Corollary 1. Let $x_j = k/n$ for $j=1,\ldots,n$ and $u_i = \sqrt{k/\pi}$ for $i=1,\ldots,n$. Then $z_D(u) \le z_{LP} \le z_{LP}(x)$ and, under condition (7),

$$z_D(u) \sim z_{LP} \qquad \text{almost surely,}$$

$$z_{LP}(x) \sim z_{LP} \qquad \text{almost surely.}$$

In addition, in [22], Papadimitriou gives a heuristic which almost surely provides a solution with value $z_H \sim z_{IP}$. The complexity of the heuristic is O(nlogn). Combining this result with the fact that $z_D(u)$ can be computed in linear time, we have a very fast procedure which will almost surely

(i) find a solution with a value close to the optimum,

(ii) prove that the value of this solution is within .3% of the optimum.

Finding the exact optimum is much more expensive as will be shown in Theorem 3. But first we give the proof of Theorem 2.

<u>Proof of Theorem 2</u>. To obtain a probabilistic upper bound on z_{LP}, we are first going to consider the LP solution

$$x_j = k/n \qquad \text{for } j=1,\dots,n \qquad .$$

and the values of y_{ij} as defined in Lemma 2. Let $d_i = \sum_{j=1}^{n} d_{ij}y_{ij}$ for $i=1,\dots,n$. We must get a probabilistic estimate of d_i for $i=1,\dots,n$. Let $\varepsilon = (\frac{k\log n}{n})^{1/3}$, $r = (\frac{1}{k\pi(1-\varepsilon)})^{1/2}$ and let S_r be the square $[r,1-r]^2$. We show first

$$(8) \qquad \Pr\left(d_i \ge \frac{2}{3\sqrt{k\pi}}(1+o(1)) \mid X_i \in S_r\right) \le 2e^{-\frac{\varepsilon^2 n}{9k}}$$

$$(9) \qquad \Pr\left(d_i \ge \frac{4}{3\sqrt{k\pi}}(1+o(1)) \mid X_i \notin S_r\right) \le 2e^{-\frac{4\varepsilon^2 n}{9k}}$$

If $X_i \in S_r$, then a circle C_i of radius r centered at X_i is entirely contained in S_0. The number N of points lying in this circle stochastically dominates the binomial $B(n, \pi r^2)$ (since $X_i \varepsilon C_i$). We define independent random variables W_j, j=1,2,...,n as follows:

Let

$$W_j = \begin{cases} d_{ij} & \text{if } X_j \varepsilon C_i \\ 0 & \text{otherwise}. \end{cases}$$

We note that $E(W_j) = 2\pi r^3/3$ $(j \neq i)$. If $N \geq \lceil\frac{n}{k}\rceil$ then $d_i \leq \frac{k}{n} \sum_{j=1}^{n} W_j$. Now, by Lemma 1,

$$\Pr(N < \lceil\frac{n}{k}\rceil) = \Pr(N \leq (1-\varepsilon)n\pi r^2) \leq e^{-\frac{\varepsilon^2}{2} n\pi r^2} .$$

Furthermore, if $\hat{W}_j = W_j/r \;\varepsilon\; [0,1]$, then by Lemma 1,

$$\Pr(\sum_{j=1}^{n} \hat{W}_j \geq (1+\varepsilon)(n-1) \frac{2\pi r^2}{3}) \leq e^{-\frac{\varepsilon^2}{3}(n-1)\frac{2\pi r^2}{3}}$$

and (8) follows.

To prove (9), we note that if $X_i \varepsilon S_0 - S_r$, we can at worst find a quadrant of a circle centered at X_i with radius 2r and contained entirely within S_0. The area of this quadrant is $\pi(2r)^2/4$ and we apply the same method as above with $E(W) = 4\pi r^3/3$.

We are now ready to bound z_{LP}.

$$z_{LP} \leq \sum_{i=1}^{n} d_i = \sum_{X_i \varepsilon S_r} d_i + \sum_{X_i \varepsilon S_0 - S_r} d_i .$$

By Lemma 1,

$$\Pr\{|X \cap S_r| \le n(1-2r)^2(1-\varepsilon)\} < e^{-\frac{\varepsilon^2}{2} n(1-2r)^2}$$

and thus

$$\Pr\{z_{LP} \ge (1+o(1))((1-2r)^2 n \frac{2}{3\sqrt{k\pi}} + (1-(1-2r)^2)n\frac{4}{3\sqrt{k\pi}})\} \le (2n+1)\ e^{-\frac{2\varepsilon^2}{9}n/k}$$

giving

(10) $$z_{LP} \le (1+o(1))\,\frac{2n}{3\sqrt{k\pi}} \quad \text{almost surely.}$$

To obtain a probabilistic lower bound on z_{LP}, we consider the dual problem (6). Let $u_i = r$ for $i=1...n$. Then by Lemma 3

(11) $$z_{LP} \ge \sum_{i=1}^{n} u_i - k \max_{j} \left(\sum_{i=1}^{n} (u_i - d_{ij})^+ \right)$$

For fixed j, consider random variables $U_i = (u_i - d_{ij})^+$.

Setting $u_i = r$ we find $E(U_i) = \frac{\pi r^3}{3}$ for $i \ne j$ and $X_j \varepsilon S_r$, whereas these values decrease for points $X_j \varepsilon S_o - S_r$. Rescaling U to [0,1] and applying Lemma 1 to $X_j \varepsilon S_r$ we find

$$\Pr\left(\sum_{i=1}^{n} U_i \ge (1+\varepsilon)\frac{n\pi r^3}{3}\right) \le e^{-\frac{\varepsilon^2}{9}n/k}$$

and thus for $k = o(\frac{n}{\log n})$ we have

$$\max_{j} \left(\sum_{i=1}^{n} U_i\right) \le (1+\varepsilon)\frac{n\pi r^3}{3} \quad \text{a.s.}$$

giving

(12) $$Z_{LP} \geq nr - (1+\varepsilon)kn\pi r^3/3 = (1-o(1))\,\frac{2n}{3\sqrt{k\pi}} \quad \text{a.s.}$$

Combining this with (10) yields the theorem. □

One might expect then that an LP-based branch and bound procedure performs well, since z_{LP} provides a good bound. However, we can prove

Theorem 3. Assume $k/\log n \to \infty$ and $n/k^2\log n \to \infty$.

Then there exists a constant $\alpha > 0$ such that a branch and bound procedure that branches by fixing a variable x_j to 0 or 1 at each node of the search tree which is not pruned and uses the LP bound to prune the search tree will almost surely explore at least $n^{\alpha k}$ nodes.

Proof: Each node of the branch and bound tree is associated with two sets J_o and J_1 where $J_t = \{j: \ x_j$ is fixed at t in the associated subproblem$\}$ for $t=0,1$. Let $z_{LP}(J_0,J_1)$ denote the LP bound computed at this node, i.e. the value of z_{LP} when we make the restriction $x_j = t$ for $j \varepsilon J_t$, $t=0,1$. We prove the theorem by showing that for some constants $\beta,\gamma > 0$ (to be determined) the following holds almost surely:

(13) For any $J_0, J_1 \subset \{1,\ldots,n\}$ such that

$$J_0 \cap J_1 = \emptyset,\ |J_0| \leq \beta n/k\log n,\ |J_1| \leq \gamma k, \text{ we have}$$

$$z_{LP}(J_0,J_1) \leq .3769\ \frac{n}{\sqrt{k}}$$

For then we almost surely have to branch at every node in which $|J_0| \leq \beta n/k\log n$ and $|J_1| \leq \gamma k$ even if we have an optimal solution of the integer program as our current best solution - by Theorem 1.

This implies that the algorithm must explore at least

$$(14) \qquad \binom{\lfloor \beta n/k\log n\rfloor + \lfloor \gamma k\rfloor}{\lfloor \gamma k\rfloor} = n^{\gamma(1-o(1))k} \text{ nodes.}$$

Since β can be chosen arbitrarily close to 1 the theorem will follow. To verify (14) imagine that setting $x_j = 0$ means branching to the left and setting $x_j = 1$ means branching to the right. (13) implies that our tree contains a copy of all possible paths which make $\lfloor \gamma k\rfloor$ right branches and $\lfloor \beta n/k\log n\rfloor$ left branches. The number of such paths is precisely the left hand side of (14).

Let F denote the family of such pairs J_0, J_1.

Thus let $J_0, J_1 \subset \{1,\ldots,n\}$ be disjoint, $\bar{J} = \{j \notin J_0 \cup J_1\}$, $\bar{n} = |\bar{J}|$, and $\bar{k} = k-|J_1|$. Consider the following solution to the associated linear program.

$$x_j = \begin{cases} 0 & \text{if } j \in J_0 \\ 1 & \text{if } j \in J_1 \\ \bar{k}/\bar{n} & \text{if } j \in \bar{J}. \end{cases}$$

The values of y_{ij} are then defined as in Lemma 2, but only using $j \in \bar{J}$ to form the sequence $j_1(i), j_2(i), \ldots, j_{\bar{n}}(i)$. This choice of y_{ij} is feasible although usually not optimum. However this is sufficient since we only need to compute an upper bound on $z_{LP}(J_0,J_1)$. We can assume w.l.o.g. that $|J_0| = \lfloor \beta n/k\log n\rfloor$ and $|J_1| = \lfloor \alpha k\rfloor$. Let $\varepsilon > 0$ be small and $r = \sqrt{\frac{1}{(1-\varepsilon)\pi\bar{k}}}$ and proceed as in the proof of Theorem 2, defining variables $W_1, W_2, \ldots W_{\bar{n}}$ for each i. We find that for $\varepsilon < \frac{1}{2}$ and n large

$$\Pr\left(z_{LP}(J_0,J_1) > \frac{2n}{3\sqrt{\pi\bar{k}}}(1 + 3\epsilon)\right) \le (2n+1)\ e^{-\frac{2\epsilon^2\bar{n}}{9\bar{k}}} .$$

Since $|F| \le n^{\beta n/k\log n + \gamma k}$ we find

$$\Pr\left(\exists\,(J_0,J_1)\epsilon F\colon z_{LP}(J_0,J_1) > \frac{2n}{3\sqrt{\pi\bar{k}}}(1 + 3\epsilon)\right) \le$$

$$(2n+1)n^{\beta n/k\log n + \gamma k}\ e^{-\frac{2\epsilon^2\bar{n}}{9\bar{k}}} .$$

Taking $\beta = \epsilon^2/5$, $\gamma = \epsilon$ and ϵ sufficiently small that $\frac{2(1+3\epsilon)}{3\sqrt{\pi(1-\epsilon)}} \le .3769$ yields

$$\max\ \{z_{LP}(J_0,J_1)\colon (J_0,J_1)\ \epsilon\ F\} \le .3769\ \frac{n}{\sqrt{k}} \text{ almost surely.}$$

Any $\alpha < \gamma$ can be used to give the theorem.

□

3. A Graphical Model

This section is concerned with the following graphical model. Let G be a random graph with n nodes, where each edge occurs independently with probability p. Let $X_1,\ldots,X_n$ be the nodes of the graph and d_{ij} the minimum number of edges on a path joining X_i to X_j for $1 \le i,j \le n$, where the minimum is taken over all paths joining X_i to X_j. Thus d_{ij} is the shortest distance between X_i and X_j, assuming that all edges have length one.

Let q=1-p and b=1/q. The main result of this section is the following theorem.

Theorem 4

(a) Consider $(1+\varepsilon)\log_b n \le k \le n$, where $\varepsilon > 0$ is fixed.

(i) If $n^{1+\delta} p \to \infty$ for all $\delta > 0$ fixed, then $z_{IP} = Z_{LP}$ almost surely.

(ii) In general, we only have $\lim_{n\to\infty} \Pr(z_{IP} = z_{LP}) = 1$

(b) Consider $2 \le k \le \log_b n$ and $p \min(1,kp) \ge \frac{\omega \log n}{n}$, where $\omega \to \infty$. Then $\frac{z_{IP} - z_{LP}}{z_{IP}} \le \frac{1}{1+e}$ almost surely. (Note that the condition in a(i) is satisfied.) In addition, if we let $kp \to \alpha$, $0 \le \alpha \le \infty$, and $p \to \beta$, $0 \le \beta < 1$, where α and β are fixed, then

$$\frac{z_{IP} - z_{LP}}{z_{IP}} \sim \frac{1 - (1-\alpha)^{+} a^{\alpha}}{1+a^{\alpha}} \text{ almost surely,}$$

where $a = e$ if $\beta=0$ and $(1-\beta)^{-1/\beta}$ if $\beta > 0$.

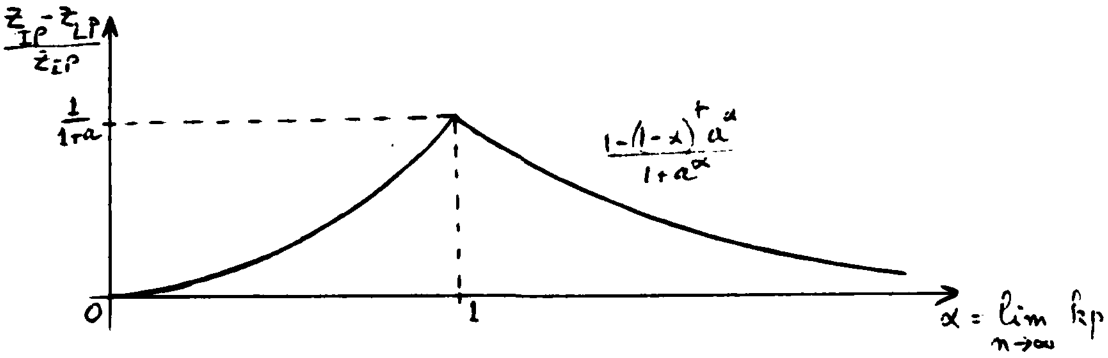


Figure 1. $\frac{z_{IP}-z_{LP}}{z_{IP}}$ as a function of kp when $2 \le k \le \log_b n$.

Proof of Theorem 4(a)

(i) This part of the theorem is a careful phrasing of a known result and is easy to prove. As $d_{ij} \ge 1$ for $i \ne j$, we must have

(15) $$z_{IP} \ge z_{LP} \ge n-k.$$

(i) follows from (15) if we can show that

$z_{IP} = n-k$ almost surely.

But $z_{IP} = n-k$ if and only if there is a set $K \subseteq X$, $|K| = k$, such that, for any $X_j \varepsilon X-K$, there exists $X_i \varepsilon K$ such that X_i and X_j are joined by an edge of $G_n(p)$, i.e., K is a <u>dominating set</u>.

Let $m = \lceil 2/\varepsilon \rceil$ and $K_i = \{X_{(i-1)k+1}, \ldots, X_{ik}\}$ for $i=1,2,\ldots m$. If none of $K_1, K_2, \ldots, K_m$ are dominating then one of the following events occurs:

$E_0 = \{ \exists\ 1 \le r \ne s \le m$ and $X_i \varepsilon K_r$ such that X_i is not adjacent in $G_n(p)$ to any vertex of $K_s\}$

$E_i = \bigcap_{i=1}^{m} F_i$ where $F_i = \{ \exists X_i \varepsilon X - \bigcup_{j=i}^{m} K_j$ such that X_i is not adjacent in $G_n(p)$ to any vertex of $K_i\}$

Now

$$\begin{aligned} Pr(E_0) &\le m^2 k\ (1-p)^k \\ &\le m^2 k n^{-(1+\varepsilon)} \\ &\le \frac{m^2 \log n}{n^{1+\varepsilon} np} && \text{as } \log_b n \le \frac{\log n}{np} \\ &= O(n^{-(1+\varepsilon/2)}) && \text{by assumption.} \end{aligned}$$

Furthermore,

$$\begin{aligned} Pr(E_1) &= \prod_{i=1}^{m} Pr(F_i) && \text{since the } F_i \text{ are independent} \\ &\le ((n-km)(1-p)^k)^m \\ &\le n^{-\varepsilon m} \\ &\le n^{-2} \end{aligned}$$

and (i) follows.

(ii) Pr(K_1 is not a dominating set)

$$\le (n-k)(1-p)^k \le n^{-\varepsilon} \to 0. \qquad \Box$$

Our proof of Theorem 4(b) will use the next two lemmas.

<u>Lemma 4</u> Consider $1 \le k \le \log_b n$. Assume $p \min(1,kp) \ge \frac{\omega \log n}{n}$, where $\omega \to \infty$. Then,

$$z_{IP} = (1+o(1))(n-k)(1+q^k) \text{ almost surely.}$$

<u>Proof</u>: For $K \subseteq X$, let N(K) be the neighbor set of K, i.e.

$$N(K) = \{X_j \in X-K: \text{ there exists an edge joining } X_j \text{ to a node of } K\}.$$

We have

$$z_{IP} \ge \min_{|K|=k} \left(|N(K)| + 2(n-k-|N(K)|)\right)$$

$$= 2(n-k) - \max_{|K|=k} |N(K)|.$$

We prove the lemma be showing that

(16) $$\max_{|K|=k} |N(K)| = (1+o(1))(n-k)(1-q^k) \text{ almost surely, and}$$

(17) $$z_{IP} = (1+o(1)) \min_{|K|=k} \left(|N(K)| + 2(n-k-|N(K)|)\right) \text{ almost surely.}$$

Consider a fixed $K \subset X$, $|K|=k$. The quantity $|N(K)|$ is distributed as $B(n-k, 1-q^k)$. Thus, by Lemma 1, for any small $\varepsilon > 0$

$$\Pr\left[|N(K)| \le (1-\varepsilon)(n-k)(1-q^k)\right] \le e^{-\frac{1}{2}\varepsilon^2(n-k)(1-q^k)} \text{ and}$$

$$\Pr\left[|N(K)| \geq (1+\varepsilon)(n-k)(1-q^k)\right] \leq e^{-\frac{1}{3}\varepsilon^2(n-k)(1-q^k)} .$$

Thus we have

$$(18) \qquad \Pr\left[\max_{|K|=k} |N(K)| \leq (1-\varepsilon)(n-k)(1-q^k)\right] \leq e^{-\frac{1}{2}\varepsilon^2(n-k)(1-q^k)}$$

$$(19) \qquad \Pr\left[\max_{|K|=k} |N(K)| \geq (1+\varepsilon)(n-k)(1-q^k)\right] \leq \binom{n}{k} e^{-\frac{1}{3}\varepsilon^2(n-k)(1-q^k)}$$

To obtain (16) we put $\varepsilon = 2(k\log\frac{n}{k} /(n-k)(1-q^k))^{\frac{1}{2}}$. We can use $\binom{n}{k} \leq (\frac{ne}{k})^k$ in (19). Then the right hand sides in (18) and (19) both → 0 sufficiently fast. Thus (16) is proved, provided that $\varepsilon < 1$.

We consider two cases. Let $0 < \alpha < 1$ be a constant.

When $kp \leq \alpha$, $q^k = (1-p)^k = [(1-p)^{1/p}]^{kp} \leq (\frac{1}{e})^{kp} \leq 1-kp + \frac{(kp)^2}{2}$ ..

So $\frac{\varepsilon^2}{4} \leq \frac{k\log n}{(n-k)kp(1 - \frac{\alpha}{2})} \to 0$ since $\log n/np \to 0$.

When $kp > \alpha$, $q^k = (1-p)^k \leq e^{-kp} \leq e^{-\alpha} < 1$. So $\frac{\varepsilon^2}{4} \leq \frac{k \log \frac{n}{k}}{(n-k)(1-e^{-\alpha})} \to 0$ since $\frac{\log x}{x} \to 0$ when $x \to \infty$.

This completes the proof of (16).

To prove (17) it suffices to show that, almost surely,

(20) every node in $X-K_1$ is joined by a path of length ≤ 2 to at least one node of K_1 where $K_1 = \{X_1,X_2,\ldots X_k\}$.

The events

$$A(j) = \{X_j \text{ is joined to } K_1 \text{ by an edge}\}$$

$$B(j) = \{X_j \text{ is joined to } K_1 \text{ via a node } X_i \neq X_j,\ X_i \notin K_1\}$$

are independent for fixed j because they have no edges in common.

$$\Pr(A(j)) = 1 - (1-p)^k = p_0, \text{ say}$$
$$\Pr(B(j)) = 1 - (1-p_0p)^{n-k-1}.$$

Hence, if N is the number of nodes not within distance 2 of K_1, then

$$\begin{aligned} \Pr(N > 0) &\leq (n-k)\ (1-p)^k\ (1-p_0p)^{n-k-1} \\ &\leq (n-k)\ (1-p_0p)^{n-1} \\ &\leq\ ne^{-(n-1)p_0p} \end{aligned}$$

If $kp \geq 1$ then $p_0 \geq 1-e^{-1}$ and so

$$\Pr(N > 0) \leq n^{-\omega/2} \qquad \text{using } p \geq \omega \log n/n.$$

If $kp < 1$ then $(1-p)^k \leq 1 - kp + \frac{k^2p^2}{2}$ and hence $p_0 \geq \frac{kp}{2}$ and then

$$\Pr\ (N > 0) \leq n^{-\omega/2} \qquad \text{using } kp^2 \geq \omega\ \log n/n.$$

This proves (20) and therefore (17) and the lemma. □

<u>Lemma 5</u> Consider $2 \leq k \leq \log_b n$. Assume $p \geq \frac{\omega\ \log n}{n}$ and $kp^2 \geq \frac{\omega}{n}$ where $\omega \to \infty$. Then

$$z_{LP} = \max(n-k,\ 2n-nkp(1+o(1))) \text{ almost surely.}$$

Proof: Given a node X_i, let $N_1(i) = \{X_j\colon d_{ij} = 1\}$ and $N_2(i) = \{X_j\colon d_{ij} = 2\}$.

First we give probabilistic estimates of $|N_1(i)|$ and $|N_2(i)|$. We will show

(21) $$\min_i |N_1(i)| = (1-o(1))np \text{ almost surely,}$$

(22) $$\max_i |N_1(i)| = (1+o(1))np \text{ almost surely, and}$$

(23) $$\min_i |N_2(i)| \geq \min\left(\frac{n}{k}, (1-o(1))nq\right) \text{ almost surely.}$$

Note that $|N_1(i)|$ is distributed as $B(n-1,p)$. So, by Lemma 1,

$$\Pr\left(\min_i |N_1(i)| \leq (1-\varepsilon)(n-1)p\right) \leq n\, e^{-\frac{1}{2}\varepsilon^2(n-1)p}$$

$$\Pr\left(\max_i |N_1(i)| \geq (1+\varepsilon)(n-1)p\right) \leq n\, e^{-\frac{1}{3}\varepsilon^2(n-1)p}.$$

Putting $\varepsilon = 3(\log n/(n-1)p)^{\frac{1}{2}}$ yields (21) and (22).

Now consider $|N_2(i)|$. We will assume $p \to 0$ (otherwise $N_1(i)$ is a dominating set by Theorem 4(a), and (23) follows). Conditional on $|N_1(i)|$, the quantity $|N_2(i)|$ is distributed as $B(n_2,p_2)$, where $n_2 = n - |N_1(i)| - 1$ and $p_2 = 1 - (1-p)^{|N_1(i)|}$. By Lemma 1,

$$\Pr\left(\min_i |N_2(i)| \leq (1-\varepsilon)n_2p_2\right) \leq n\, e^{-\frac{1}{2}\varepsilon^2 n_2p_2}.$$

Set $\varepsilon = 3(\log n/n_2p_2)^{\frac{1}{2}}$. We have to show $\varepsilon < 1$. Note that $n_2 = (1-o(1))n$ and $p_2 = 1 - (1-p)^{(1+o(1))np} \geq 1-e^{-(1+o(1))np^2}$ almost surely.

If $np^2 \geq \delta > 0$ where δ is fixed, then

$$\frac{\varepsilon^2}{4} \leq \frac{\log n}{(1+o(1))n(1-e^{-\delta})} \to 0.$$

If $np^2 = o(1)$, then

$$\frac{\varepsilon^2}{4} - \frac{\log n}{n^2 p^2} = \frac{1}{\log n} \left(\frac{\log n}{np} \right)^2 \to 0.$$

So we have just shown that, almost surely,

$$\min_i |N_2(i)| \geq (1-o(1))n_2 p_2.$$

Next we will use the fact that $kp^2 \geq \frac{\omega}{n}$ to show $n_2 p_2 \geq \frac{n}{k}$ almost surely.

If $np^2 \geq \delta$, $0 < \delta < 1$ fixed, then almost surely

$$n_2 p_2 \geq (1+o(1))n(1-e^{-\delta}) \geq \frac{n}{k} \text{ for } k \geq 2 \text{ and } \delta \text{ close enough to } 1.$$

If $np^2 \leq \delta < 1$, then $1 - e^{-(1+o(1))np^2} \geq np^2(1 - \frac{np^2}{2})$. So

$$n_2 p_2 \geq (1+o(1))n^2 p^2 (1 - \frac{\delta}{2}) \geq (1+o(1)) \frac{n\omega}{k} (1 - \frac{\delta}{2}) \geq \frac{n}{k} \text{ almost surely.}$$

This complete the proof of (23).

Now we are ready to get a probabilistic estimate of z_{LP}. First we obtain an upper bound by considering the solution

(24) $\qquad x_j = \frac{k}{n}$ for $j=1,\ldots,n$ and y_{ij} defined in Lemma 2.

Let $\delta = \min_i |N_1(i)|$ be the minimum degree of $G_n(p)$. Note that, if $\delta \geq \frac{n}{k} - 1$, then $z_{LP} = n-k$. For, using the solution (24), we have $d_i = \sum_{j=1}^{n} d_{ij}y_{ij} = 1 - \frac{k}{n}$ for $i=1,\dots,n$. On the other hand, if $\delta < \frac{n}{k} - 1$, then $d_i \leq \frac{k}{n}\delta + 2\frac{k}{n}(\frac{n}{k} - 1 - \delta)$. ($y_{ij}$ only takes positive values for points X_j at distance one or two of X_i since, by (23), the number of points at distance 2 is at least $\min(\frac{n}{k}, (1-o(1))nq)$ which is more than the $\frac{n}{k} - 1 - \delta$ points needed.) Therefore $z_{LP} \leq n\sum_{i=1}^{n} d_i \leq 2n-k\delta$, almost surely.

To obtain a probabilistic lower bound for z_{LP} we consider the dual bound given by Lemma 3. We put $u_i = 2 - \frac{1}{n}$ for $i=1,\dots,n$ and let Δ denote the maximum degree of $G_n(p)$. Then

$$z_{LP} \geq n(2 - \frac{1}{n}) - k\Delta(1 - \frac{1}{n}) = 2n - (1+o(1))nkp \quad \text{almost surely.}$$

This completes the proof of Lemma 5. □

Proof of Theorem 4(b)

It follows from Lemmas 4 and 5 that

$$\frac{z_{IP} - z_{LP}}{z_{IP}} \sim \frac{(1+q^k) - \max(1,2-kp)}{1+q^k} \quad \text{almost surely}$$

$$= \frac{q^k - (1-kp)^+}{q^k + 1} .$$

Setting $a = (1-p)^{-1/p}$ and $kp = \alpha$, we get

$$\frac{z_{IP} - z_{LP}}{z_{IP}} \sim \frac{1 - (1-\alpha)^+ a^\alpha}{1 + a^\alpha} \quad \text{almost surely.}$$

It is easy to check that the maximum of this function is achieved when $p \to 0$

and $\alpha = 1$. Then its value is $\frac{1}{1+e}$. ⌟

An interesting range of parameters which is not considered in Theorem 4 is the case $2 \le k \le \log_b n$ and $p \ge \frac{\omega \log n}{n} \ge kp^2$ where $\omega \to \infty$. In this range, the expressions for z_{IP} and z_{LP} are more complicated than those found in Lemmas 4 and 5. However we conjecture that $\frac{z_{IP} - z_{LP}}{z_{LP}} \to 0$ almost surely.

In the range covered by Theorem 4, it is easy to identify conditions under which the ratio $\frac{z_{IP} - z_{LP}}{z_{LP}}$ is almost surely bounded away from 0. For example, consider

(25) $\varepsilon \le kp \le 1/\varepsilon$, $k \ge 2$ and

(26) $(\omega \log n/n)^{1/2} \le p \le 1-\varepsilon$

where $\omega \to \infty$ and $0 < \varepsilon < 1$ is fixed.

Then $k \log b = kp\,(1 + \frac{p}{2} + \frac{p^2}{3} + \ldots) \le \frac{kp}{1-p} \le \frac{1}{\varepsilon^2}$. So $k \le \log_b n$ for n large enough and, by Theorem 4(b), there is a fixed value $f(\varepsilon) > 0$ such that

(27) $$\frac{z_{IP} - z_{LP}}{z_{IP}} \ge f(\varepsilon) \quad \text{almost surely.}$$

In addition, we can show that, under these conditions, a branch and bound algorithm based on the LP bound z_{LP} almost surely requires close to complete enumeration.

Theorem 5 Assume (25) and (26). A branch and bound procedure that branches by fixing a variable x_j to 0 or 1 at each node of the search tree which is not pruned, and uses the LP bound to prune the search tree, will almost surely expand at least $n^{(1-o(1))(k-2)}$ nodes. (The number of feasible solutions of the k-median problem is $\binom{n}{k} = n^{(1-o(1))k}$.

Proof: We first note that, under the above assumptions, $\varepsilon \le k \log b \le \frac{1}{\varepsilon^2}$ and therefore

(28) $$e^{-1/\varepsilon^2} \le q^k \le e^{-\varepsilon}.$$

In addition, the assumptions of Lemma 4 hold and $k = o(n^{1/2})$ so that

(29) $$z_{IP} \geq (1-o(1))\, n(1+q^{k}) \text{ almost surely.}$$

Let $z_{LP}(J_0,J_1)$ be the LP value of the subproblem where $J_0 = \{j: x_j$ is fixed to 0$\}$ and $J_1 = \{j: x_j$ is fixed to 1$\}$.

Let $\alpha<1$ and $\beta>0$ be fixed. We prove the theorem by showing that, for β chosen small enough, the following property holds almost surely.

(30) For any $J_0, J_1 \subseteq \{1,2,\ldots,n\}$ such that $J_0 \cap J_1 = \emptyset$, $|J_0| \leq \lceil \beta n \rceil$ and $|J_1| \leq \lceil \alpha k \rceil$,

(31) $$z_{LP}(J_0,J_1) < z_{IP}.$$

This implies that the algorithm must explore at least

(32) $$\binom{\lceil \beta n\rceil + \lceil \alpha k\rceil}{\lceil \alpha k\rceil} \geq \left(\frac{\beta n}{\alpha k}\right)^{\alpha k} = n^{(1-o(1))\alpha k} \text{ nodes.}$$

To verify (32), imagine that setting $x_j = 0$ means branching to the left and setting $x_j = 1$ means branching to the right. (30) - (31) imply that any tree contains all possible paths which make $\lceil \alpha k \rceil$ right branches and $\lceil \beta n \rceil$ left branches. The number of such paths is precisely the left hand side of (32).

We now turn to the proof of (31). As increasing J_0 or J_1 only serves to increase z_{LP} we can restrict our attention to $|J_0| = \lceil \beta n \rceil$ and $|J_1| = \lceil \alpha k \rceil$.

Using Lemma 1 we can easily prove that the following holds almost surely for $G_n(p)$:

(33) $J \subseteq \{1,2,\ldots,n\}$ and $|J| = \lceil \alpha k \rceil$ implies

$$|N(J)| \geq (1-o(1))n(1-q^{\alpha k}) \qquad \text{(see (18))}$$

Furthermore, it is easy to see that

(34) $$\text{diam}\,(G_n(p)) = 2 \text{ almost surely.}$$

where diam refers to the diameter of $G_n(p)$.

Indeed Pr(there exists $i,j \in \{1,2,\ldots,n\}$ such that i,j are not joined by a path of length 2)

$$\leq \binom{n}{2} (1 - p^2)^{n-2}$$
$$\leq n^2 e^{-(n-2)p^2}$$
$$\leq n^2 e^{-(\omega \log n)(n-2)/n} \to 0.$$

Thus (34) is proved $(\Pr[\mathrm{diam}(G_n(p)) = 1] = p^{\binom{n}{2}} \to 0)$ To obtain an upper bound on $z_{LP}(J_0,J_1)$ let

$$x_j = \begin{cases} 0 & \text{if } j \epsilon J_0 \\ 1 & \text{if } j \epsilon J_1 \\ \gamma & \text{if } j \notin J_0 \cup J_1 \end{cases}$$

where $\gamma = (k - \lceil \alpha k \rceil)/(n - \lceil \beta n \rceil - \lceil \alpha k \rceil)$.

The values for y_{ij} are then chosen as follows:

$i \epsilon J_1 \quad : \quad y_{ii} = 1 \quad$ and $y_{ij} = 0 \quad j \neq i$

$i \epsilon N(J_1) : \quad y_{it} = 1 \quad$ and $y_{ij} = 0 \quad j \neq t$

where t is a node of $J_1 \cap N(i)$.

$i \notin J_1 \cup N(J_1)$: the values are as defined in Lemma 2.

With this solution we find, using (34) that

$$\begin{aligned} d_i &= 0 && \text{if } i \epsilon J_1 \\ &= 1 && \text{if } i \epsilon N(J_1) \\ &\leq \gamma(\delta - s_i) + 2(1 - \gamma(\delta - s_i)) && \text{if } i \notin J_1 \cup N(J_1) \end{aligned}$$

where $s_i = |N(i) \cap J_o|$, δ is the minimum node degree and Δ is the maximum node degree in $G_n(p)$.

To compute an upper bound on z_{LP}, we will distinguish between the cases $\gamma\delta \leq 1$ and $\gamma\delta > 1$.

First assume that $\gamma\delta > 1$. We use the bound

$$d_i \leq \gamma(\delta - s_i) + 2(1 - \gamma(\delta - s_i)) \leq 1 + \gamma s_i.$$

$$z_{LP} \leq |N(J_1)| + \sum_{i \notin J_1 \cup N(J_1)} (1 + \gamma s_i)$$
$$\leq |N(J_1)| + n - |N(J_1)| + \gamma\Delta |J_o|$$

$$= n + \frac{\beta(k-\lceil \alpha k \rceil)p}{1-\beta} n + o(n).$$

Since kp is bounded above by a constant as a consequence of (25), we simply choose β small enough to get our bound on z_{LP}. Then (31) follows from (28) and (29).

Now assume that $\gamma\delta \le 1$. We use the bound

$$d_i \le \gamma(\delta - s_i) + 2(1-\gamma(\delta-s_i)) = 2 - \gamma\delta + \gamma s_i.$$

$$z_{LP} \le |N(J_1)| + \sum_{i \notin J_1 \cup N(J_1)} (2 - \gamma\delta + \gamma s_i)$$

$$= |N(J_1)| + (2 - \gamma\delta)(n - |N(J_1)|) + \gamma\Delta|J_o|$$

$$= (2-\gamma\delta)n - (1-\gamma\delta)|N(J_1)| + \gamma\Delta|J_o|$$

$$\le \left[1 + q^{\lceil \alpha k \rceil}\left(1 - \frac{k-\lceil \alpha k \rceil}{1-\beta} p\right) + \frac{\beta(k-\lceil \alpha k \rceil)}{1-\beta} p\right]n + o(n)$$

where the last inequality follows from the relations

$$|N(J_1)| \ge (1-o(1))n(1-q^{\lceil \alpha k \rceil})$$

$$\gamma\delta = (1+o(1)) \frac{k-\lceil \alpha k \rceil}{1-\beta} p$$

$$\Delta = (1+o(1))np.$$

Therefore

$$z_{IP} - z_{LP} \ge \left[q^{\lceil \alpha k \rceil}((1-p)^m - 1 + mp) - \frac{\beta mp}{1-\beta}(1-q^{\lceil \alpha k \rceil})\right]n + o(n)$$

where $m = k - \lceil \alpha k \rceil$. Note that $\gamma\delta \sim \frac{mp}{1-\beta} \le 1$ implies $mp < 1$.

Next we show that $S_m = (1-p)^m - 1 + mp$ is bounded below by a positive constant. This will imply that $z_{IP} - z_{LP} > 0$ by choosing β small enough.

We assume that α is chosen so that $\alpha \le 1 - \frac{2}{k}$. This implies $m \ge 2$. Now

$$\begin{aligned} S_m &= S_{m-1} + p(1-(1-p)^{m-1}) \\ &\geq S_{m-1} + p(1-e^{-(m-1)p}) \\ &\geq S_2 + p \sum_{i=3}^{m-1} (1-e^{-(i-1)p}) \\ &\geq p^2 + p(m-\lfloor m/2 \rfloor)(1-e^{(\lfloor m/2 \rfloor -1)p}). \end{aligned}$$

If k is fixed, then p is bounded below by a constant as a consequence of (25). Therefore S is bounded below by a constant.

If $k \to \infty$, then $m \sim (1-\alpha)k$. Thus mp and hence S_m is bounded below by a constant using (25).

This completes the proof of (31). Note that (32) and the condition $\alpha \leq 1-\frac{2}{k}$ imply the bound $n^{(1-o(1))(k-2)}$ announced in the statement of the theorem. □

In [4], a different graphical model is associated with the variation of the k-median problem known as the k-plant location problem. The k-plant location problem is defined using two sets $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$. The quantity d_{ij} is defined for each $1 \leq i \leq m$ and $1 \leq j \leq n$. The problem consists of finding a set $S \subseteq X$, $|S| = k$, that minimizes $\sum_{i=1}^{m} \min_{j \in S} d_{ij}$.

A k-plant location problem arises from a graph G by defining X as its node set, Y as its edge set and $d_{ij} = 0$ if X_j is incident with Y_i, 1 otherwise. (The problem is to find k nodes that cover the maximum number of edges of G.) It is shown in [4] that

$$z_{IP} = z_{LP} \qquad\qquad \text{almost surely}$$

when $G = G_n(p)$ is a random graph with $0<\varepsilon\leq p\leq 1-\varepsilon$, ε fixed, and $k \leq n^{\alpha}$, $\alpha<1/6$ fixed.

4. A Tree Model

This section is concerned with the following tree based model: we are given a random tree T_n with node set $X = \{X_1, \ldots, X_n\}$ where each of the n^{n-2} different trees is equally likely to occur. The distance d_{ij} is the number of edges in the unique path from X_i to X_j in T_n. This section contains a probabilistic result (Theorem 7) and a deterministic one (Theorem 6).

Kolen[13] proved that $z_{IP} = z_{LP}$ for every SPLP defined on a tree. For the k-median problem, this equality does not always hold as shown in Theorem 6. In fact we show in Theorem 7 that, for random trees on n nodes, the number of values of k such that $z_{LP} \neq z_{IP}$ is almost surely at least cn, for some constant $c > 0$.

Theorem 6

(a) For $k = 1$ or $k \geq \lfloor\frac{n-1}{2}\rfloor$, $z_{IP} = z_{LP}$ for every tree on n nodes.

(b) For $2 \leq k < \lfloor\frac{n-1}{2}\rfloor$, and $n \neq 8$, there is a tree on n nodes such that $z_{IP} \neq z_{LP}$.

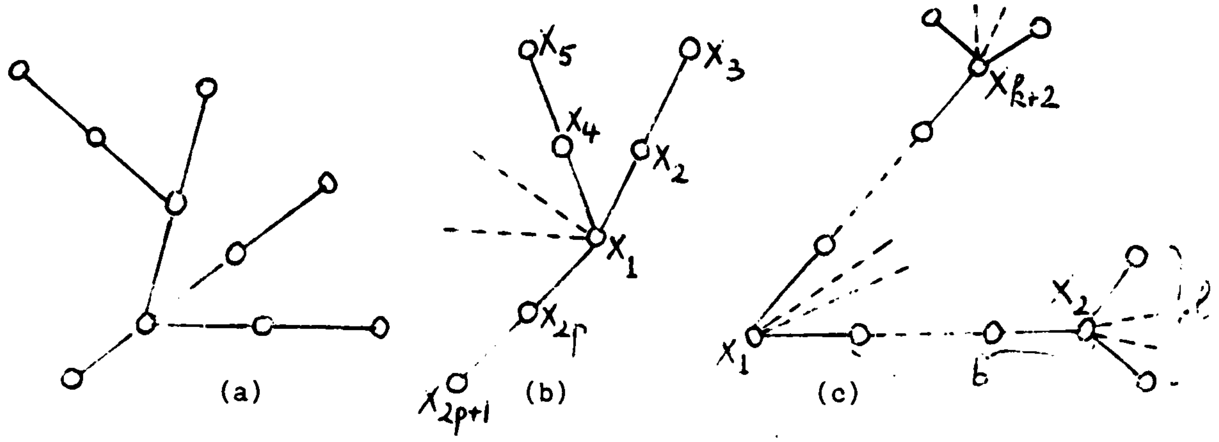
(c) There is an infinite family of trees such that $\frac{z_{IP} - z_{LP}}{z_{IP}} \to \frac{k-1}{2k}$.

It would be interesting to perform a worst-case analysis of the k-median problem and its LP relaxation on trees. We conjecture that the ratio $\frac{k-1}{2k}$ found in (c) is the worst-case bound.

Proof of Theorem 6: For the 1-median problem, it is well-known that $z_{IP} = z_{LP}$ for every choice of d_{ij}, $1 \leq i,j \leq n$. For example, this result appears in Mukendi [18].

When $k \geq \lfloor\frac{n}{2}\rfloor$, $z_{IP} = z_{LP} = n - k$ follows from the fact that every tree on n nodes has a dominating set of cardinality at most $\lfloor\frac{n}{2}\rfloor$. (A tree is bipartite and a color class dominates it).

To complete the proof of Theorem 6(a), it suffices to consider the case where n is even and $k = \frac{n}{2} - 1$. The only trees which do not have a dominating set of size k are constructed inductively from a path with 4 nodes by adding paths $P_i = (v_1{}^i, v_2{}^i, v_3{}^i)$ where $v_1{}^i$ is one of the nonleaf nodes of the current tree and $v_2{}^i$, $v_3{}^i$ are two new nodes. (See Figure 2(a)). From the construction $z_{IP} = n-k+1 = \frac{n}{2}+ 2$. Using the dual values $u_j = 2$ if X_j is a leaf, 1 if not, Lemma 3 yields $z_{LP} \geq \frac{n}{2} + 2$. Therefore $z_{IP} = z_{LP}$.



(a) (b) (c)

Figure 2

To prove Theorem 6(b) when n is odd, consider the tree of Figure 2(b). Let $p = \frac{n-1}{2}$. An optimal solution of the k-median problem is to take $S = \{X_1, X_2, X_4, X_6, \ldots, X_{2(k-1)}\}$. Then $z_{IP} = 3p - 2(k-1)$. We get a feasible solution of the LP relaxation by setting $x_1 = \frac{p-k}{p-1}$ and $x_{2i} = \frac{k-1}{p-1}$ for $i = 1, \ldots, p$. This yields

$$z_{LP} \leq (3p^2 - 2pk - p + k - 1) / (p-1).$$

Therefore $z_{IP} - z_{LP} \geq \frac{k-1}{p-1} > 0$.

To prove Theorem 6(b) when n is even, $n \neq 8$, we first consider the case $k \geq 3$. Add a node X_{2p+2} adjacent to X_{2p} to the tree of Figure 2(b). Then it is optimum to choose X_{2p} in S and we can also choose $x_{2p} = 1$ in the LP solution. Removing X_{2p}, X_{2p+1} and X_{2p+2}, we are back to the case where n is odd and $k \geq 2$. Now consider the case $n \geq 10$ even and $k = 2$. Add three nodes to the graph of Figure 2(b), namely X_{2p+1+i} adjacent to X_{2i} for $i = 1,2,3$. Then $z_{IP} = 3p+3$ but there is a better LP solution, namely $x_1 = 1$ and

$x_2 = x_4 = x_6 = 1/3$. This yields $z_{LP} = 3p + 1$.

Finally, to prove Theorem 6(c), consider the tree of Figure 2(c). The node X_1 has degree k+1 in the tree. Each branch incident with X_1 contains b nonleaf nodes and ℓ leaf nodes where $b \to \infty$, $\ell \to \infty$ and ℓ grows much faster than b. We denote by $X_2, \ldots, X_{k+2}$ the (k+1) nodes of the tree which are incident with leaves. Then, an optimal solution of the k-median problem is $\{X_1, X_2, X_3, \ldots, X_k\}$.

$$z_{IP} = (k-1)\ell + 2(b+1)\ell + O(kb^2)$$

where the last term accounts for all the nonleaf nodes. Ignoring the lower order terms,

$$z_{IP} \sim 2b\ell.$$

To get an optimal LP solution, set $x_1 = \frac{1}{k}$ and $x_j = \frac{k-1}{k}$ for $j = 2, \ldots, k+2$.

$$z_{LP} = (k+1)\ell \times \frac{k-1}{k} + (k+1)\ell \times (b+1)\frac{1}{k} + O(kb^2), \text{ i.e.}$$

$$z_{LP} \sim \frac{k+1}{k} b\ell.$$

Therefore $\dfrac{z_{IP}-z_{LP}}{z_{IP}} \sim \dfrac{2-\frac{k+1}{k}}{2} = \dfrac{k-1}{2k}$.

□

In the next theorem we consider all the k-median problems defined on a tree, namely all $1 \leq k \leq n$ where n is the number of nodes in the tree.

<u>Theorem 7</u> Let T_n be a random tree. There exists a positive constant c such that, almost surely, $z_{IP} \neq z_{LP}$ for at least cn different values of k.

<u>Proof</u> = Consider a random tree $T_n = (V_n, E_n)$ and a <u>fixed</u> tree $T = (V,E)$. Let $v \in V$. We say that T_n contains a copy of T suspended at v if there exists $V' \subseteq V_n$ such that

(35) $T_n(V')$ is isomorphic to T under a mapping $\phi : V \to V'$.

(36) there is a unique edge of T_n with exactly one end, say v', in V' and, in addition, $v' = \phi(v)$.

Let $m = |V|$ and a = the number of automorphisms of T. Then m!/a is the number of distinct labeled graphs on m nodes which are isomorphic copies of T. We first prove that almost surely T_n contains at least $(1-o(1))(n/e^m a)$ copies of T suspended at v.

For each $V' \subseteq V_n$, $|V'| = m$, let

$$\delta(V') = \begin{cases} 1 & \text{if (35) and (36) hold,} \\ 0 & \text{otherwise.} \end{cases}$$

We note that, if $V' \cap V'' \neq \emptyset$, then $\delta(V')\delta(V'') = 0$.

Let $N = \sum_{\substack{V' \subset V_n \\ |V'|=m}} \delta(V')$

= the number of copies of T suspended at v contained in T_n.

Now for a fixed copy of T on a set of m nodes, there are $(n-m)^{n-m-1}$ ways of choosing a tree on the remaining n-m nodes and then joining it to v. Thus

$$\mathbf{E}(N) = \binom{n}{m} (m!/a) (n-m)^{n-m-1}/ n^{n-2}$$
$$\sim n/e^m a.$$

Using the Markov inequality $\Pr(Y \geq \alpha) \leq \frac{E(Y)}{\alpha}$ with $Y = (N-E(N))^4$, $E(Y) = \mu_4$ and $\alpha = \lambda^4 \mu_4$ we get the Pearson extension of the Chebychev inequality.

$$\Pr\{|N-E(N)| \geq \lambda \mu_4^{\frac{1}{4}}\} \leq \frac{1}{\lambda^4}$$

In terms of factorial moments μ_4 is given by

$$\begin{aligned} \mu_4 \quad &= \mu_{[4]} - 4\mu_{[1]}\,\mu_{[3]} + 6\mu_{[1]}^2\mu_{[2]} - 3\mu_{[1]}^4 \\ &+ 6\mu_{[3]} - 12\mu_{[1]}\mu_{[2]} + 6\mu_{[1]}^3 \\ &+ 7\mu_{[2]} - 4\mu_{[1]}^2 \\ &+ \mu_{[1]} \end{aligned}$$

where $\mu_{[i]}$ is the ith factorial moment.

$$\mu_{[i]} = \frac{n!}{(m!)^i(n-im)!} \left(\frac{m!}{a}\right)^i \frac{(n-im)^{n-im-2}}{n^{n-2}} (n-im)^i$$

We find

$$\mu_{[2]} = \mu_{[1]}^2 \, e^{-2m/n + O(\frac{1}{n^2})}$$

$$\mu_{[3]} = \mu_{[1]}^3 \, e^{-6m/n + O(\frac{1}{n^2})}$$

$$\mu_{[4]} = \mu_{[1]}^4 \, e^{-12m/n + O(\frac{1}{n^2})} .$$

In the expression for μ_4 above, the first row is the powers of n^4. When we evaluate this row we find that terms in 1 and 1/n of the exponentials disappear simultaneously, leaving a term in $\mu_{[1]}^4 \, O(\frac{1}{n^2})$, i.e. $O(n^2)$. Similarly in the next row (powers of n^3) the terms in 1 of the exponentials disappear simultaneously leaving $\mu_{[1]}^3 \, O(\frac{1}{n})$, i.e. $O(n^2)$. The last two rows are $O(n^2)$.

Thus $\mu_4 = O(n^2)$ and setting $\lambda = n^{\frac{1}{4}+4\varepsilon}$ gives

$$\Pr(|N - (1+\varepsilon)\frac{n}{e^{m}a}| \geq n^{3/4}) = O(\frac{1}{n^{1+\varepsilon}}).$$

Now we consider the fixed tree T given in Figure 3.

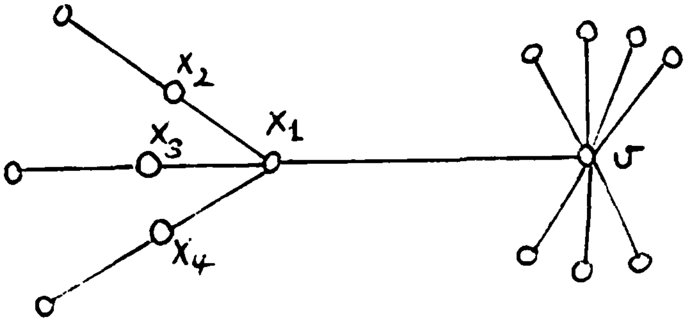


Figure 3.

Let S_k be an optimal k-median solution in T_n. We will let k increase from 1 to n. Consider any copy of T suspended at v contained in T_n, say (V',E'). Note that, if $|V' \cap S_k| \geq 1$, then $v \in S_k$. This implies that there exists a K such that for $k \geq K$, $|V' \cap S_k|$ is a nondecreasing function of k which goes from 1 to 15 (= m).

Let $z_{IP}(V') = \sum_{i \in V'} \min_{j \in S_k} d_{ij}$. When $|V' \cap S_k| = 3$, an optimal set $V' \cap S_k$ is $\{v, X_1, X_2\}$ with $z_{IP}(V') = 14$. However, consider the fractional solution $x_1 = x_2 = x_3 = x_4 = \frac{1}{2}$, $x_j = 1$ for the variable associated with node v, and $x_j = 0$ for the other nodes of V'. Let y_{ij} be defined as in Lemma 2 and $z_{LP}(V') = \sum_{i \in V'} a_{ij}y_{ij}$. The above fractional solution yields $z_{LP}(V') = 13.5$ and therefore, when $|V' \cap S_k| = 3$, $z_{IP} > z_{LP}$.

Since T_n contains almost surely at least (1-o(1))n/m!a copies of T suspended at v, there are at least as many values of k for which $z_{IP} > z_{LP}$.

□

5. The uniform cost model.

In this section, we look briefly at the model where the d_{ij}'s are drawn independently from the [0,1] uniform distribution, $1 \leq i,j \leq n$.

Here we do not assume $d_{ii} = 0$, $d_{ij} = d_{ji}$ or $d_{ij} \leq d_{ik} + d_{kj}$, as we did in the other models. The quantity d_{ij} is interpreted as the cost of assigning X_i to X_j.

The main result of this section states that, when $k \geq n(e-1)/e$, then $z_{IP} = z_{LP}$ almost surely, and when $k = o(n/\log n)$, then $\frac{z_{IP}-z_{LP}}{z_{IP}} \sim \frac{k-1}{2k}$ almost surely. The analysis is made possible by the fact that, in those ranges, the k-median problem is almost surely trivial to solve exactly or approximately. (When $k \geq n(e-1)/e$ there is an obvious optimal solution, and when $k = o(n\log n)$ every solution is close to optimum.)

Theorem 8

(a) Suppose $k = o(n/\log n)$. Then

$$z_{IP} \sim n/(k+1) \qquad \text{almost surely}$$

$$z_{LP} \sim n/2k \qquad \text{almost surely.}$$

(b) Suppose $k \geq (1+o(1))n(e-1)/e$. Then

$$z_{IP} = z_{LP} \qquad \text{almost surely.}$$

Proof = Let S be a fixed set of size k. If we take $x_j = 1$ for $j \epsilon S$ as our solution to the integer program, then the $d_i = \min_{j \epsilon S} d_{ij}$ are independently distributed as the minimum of k uniform [0,1] random variables, i.e.

$$\Pr(d_i \geq a) = a^k \qquad \text{for } 0 \leq a \leq 1,$$

and hence $\mathcal{E}(d_i) = 1/(k+1)$ for $i = 1, \ldots, n$. We first consider $k = O(n^{1/5})$. Applying Lemma 1 to $D = d_1 + \ldots + d_n$, we have

$$\Pr(|D-n/(k+1)| \geq \epsilon n/(k+1)) \leq 2e^{-\epsilon^2 n/3(k+1)} .$$

Now put $\epsilon = n^{-1/5}$. In addition, $\binom{n}{k}e^{-\epsilon^2 n/3(k+1)} = o(e^{-n^{1/3}})$, so $z_{IP} \sim n/k+1$ almost surely.

Assume now that $k/n^{1/6} \to \infty$. Then $\omega = (\frac{n}{k \log n})^{\frac{1}{4}} \to \infty$. Set

$$\delta_i = \begin{cases} d_i & \text{if } d_i \leq \omega/k \\ 0 & \text{otherwise} \end{cases}$$

and note that $E(\delta_i) = \frac{1}{k+1}\,(1-(1-\frac{\omega}{k})^k(\omega-1)) = \frac{1-o(1)}{k+1}$. We rescale the δ_i to [0,1] and apply Lemma 1.

$$\Pr\Big(\sum_{i=1}^{n} \delta_i/(\omega/k) \leq (1-\tfrac{1}{\omega})\; n\; E(\delta_i/(\omega/k))\Big) \leq e^{-\frac{n}{2\omega^3}(1-o(1))} .$$

As $\delta_i \leq d_i$ we deduce

$$\Pr(z_{IP} \leq n\,\frac{1-o(1)}{k+1}) \leq \binom{n}{k}\; e^{-\frac{n}{2\omega^3}(1-o(1))}$$

and hence if $k = o(\frac{n}{\log n})$

$$z_{IP} \geq \frac{(1-o(1))n}{k+1} \quad \text{almost surely.}$$

On the other hand, taking $S = \{1,2,\ldots k\}$ we can show easily that

$$z_{IP} \leq \frac{(1+o(1))n}{k+1} \quad \text{almost surely.}$$

Now we prove the second part of Theorem 8(a). We put $x_j = k/n$ as usual. Then d_i is dominated probabilistically by k/n times the sum of the $\lceil n/k \rceil$

smallest out of n independent [0,1] uniform random variables. Thus

$$\mathbb{E}(d_i) \le \quad (k/n) \sum_{t=1}^{\lceil n/k \rceil} t/(n+1) \quad = (1+o(1))/2k.$$

Applying Lemma 1 in the usual way shows that

$$z_{LP} \le (1+o(1))n/2k \qquad \text{almost surely.}$$

On the other hand, consider the dual solution $u_i = 1/k$ for $i = 1, \ldots, n$. Then, by Lemma 3,

$$z_{LP} \ge \sum_{i=1}^{n} u_i - k \max_{j=1,\ldots n} \left(\sum_{i=1}^{n} (u_i - d_{ij})^+ \right).$$

As in Theorem 2, for fixed j, we consider random variables $U_i = (u_i - d_{ij})^+$. Setting $u_i = \frac{1}{k}$ we find $E(U_i) = \frac{1}{2k^2}$. Rescaling the U_i to [0,1] and applying Lemma 1

$$\Pr\left(\sum_{i=1}^{n} k\, U_i \ge (1+\varepsilon) \frac{n}{2k} \right) \le e^{-\frac{\varepsilon^2}{3} \frac{n}{2k}}.$$

Hence, if $n/k = \theta^3 \log n$ where $\theta \to \infty$, then taking $\varepsilon = 1/\theta$ yields

$$z_{LP} \ge \frac{n}{k} - k(1+\varepsilon) \frac{n}{2k^2} \qquad \text{almost surely}$$

$$= (1-o(1)) \frac{n}{2k}.$$

This completes the proof of Theorem 8(a).

Now consider the case where k is sufficiently large so that each point X_i can be assigned to the cheapest point $X_{j(i)}$, defined by $d_{ij(i)} = \min_{j=i,\ldots,n} d_{ij}$. Then clearly $z_{IP} = z_{LP}$.

For $j = 1,\ldots,n$, let N_j be the number of points X_i assigned to X_j according to the above scheme. N_j is asymptatically distributed according to a Poisson process with mean 1; in particular $\Pr(N_j=0) \sim 1/e$. Therefore $E(|\{j : N_j=0\}|) \sim n/e$. To show $Z = |\{j: N_j = 0\}| \leq (1+o(1))\frac{n}{e}$ almost surely we use the generalized Markov inequality $\Pr(Z \geq a) \leq \frac{E(\phi(Z))}{\phi(a)}$ for any non-negative monotone increasing ϕ. We let $k = \lceil n^{1/3} \rceil$, $a = (1+\varepsilon)\frac{n}{e}$ where $\varepsilon = n^{-1/4}$ and $\phi(a) = \max\{0,a(a-1)\ldots(a-k+1)\}/k!$ and note that $\phi(Z)$ = the number of k-sets S for which $N_j = 0$, $j \varepsilon S$. This gives

$$\Pr(Z \geq (1+\varepsilon)\tfrac{n}{e}) \leq \frac{\binom{n}{k}(1-\frac{k}{n})^n}{a(a-1)\ldots(a-k+1)/k!}$$

$$= 0\left((1+\varepsilon)^{-k}\right).$$

This completes the proof of Theorem 8. □

As for the Euclidean and graphical models, we can show

<u>Theorem 9</u> Suppose $k = o((n/\log n)^{1/2})$ and $k \to \infty$. Then an LP based branch and bound algorithm almost surely explores at least $n^{(1-o(1))k}$ nodes of the search tree.

<u>Proof</u> = Let $z_{LP}(J_0,J_1)$ be the LP value of the subproblem where $J_0 = \{j : x_j \text{ is fixed to } 0\}$ and $J_1 = \{j : x_j \text{ is fixed to } 1\}$. Assume that $\alpha < 1$ is close to 1, that β is large and that α and β have been chosen so that αk and βk are integer. To prove the theorem, it suffices to show that, almost surely,

(37) for any J_0, $J_1 \subseteq \{1,2,\ldots,n\}$ such that $J_0 \cap J_1 = \emptyset$, $|J_1| \leq \alpha k$ and $|J_0| \leq n-(\beta-\alpha)k$, we have $z_{LP}(J_0,J_1) < z_{IP}$.

As increasing J_0 or J_1 only serves to increase z_{LP}, we can restrict our attention to $|J_1| = \alpha k$ and $|J_o| = n-(\beta-\alpha)k$. Let $L = \alpha k$, $K = \beta k$ and $\gamma = \frac{(1-\alpha)k}{n-|J_o|-|J_1|} = \frac{1-\alpha}{\beta} = \frac{k-L}{K}$. To obtain an upper bound on $z_{LP}(J_o,J_1)$, let

$$x_j = \begin{cases} 0 & \text{if } j\varepsilon J_o \\ 1 & \text{if } j\varepsilon J_1 \\ \gamma & \text{if } j\notin J_o \cup J_1. \end{cases}$$

Consider a fixed i and suppose $c_{ij_1} \leq c_{ij_2} \leq \ldots \leq c_{ij_n}$. Let $t = \min\{s : j_s \varepsilon J_1\}$. Let y_{ij} be given by Lemma 2 and $d_i = \sum_{j=1}^{n} d_{ij}y_{ij}$. The expected value of d_i, conditional on knowing the value of t, is

$$(38) \qquad \mathrm{Exp}(d_i|t) = \gamma \sum_{i=1}^{t} \frac{i}{K+1} + (1-\gamma t)\frac{t}{K+1}$$

$$= \frac{t}{K+1} - \frac{\gamma(t-1)t}{2(K+1)} \qquad \text{if } t \leq \lfloor\gamma^{-1}\rfloor,$$

and

$$(39) \qquad \mathrm{Exp}(d_i|t) = \gamma \sum_{i=1}^{\lfloor\gamma^{-1}\rfloor} \frac{i}{K+1} + \gamma(\gamma^{-1}-\lfloor\gamma^{-1}\rfloor)\frac{\lfloor\gamma^{-1}\rfloor+1}{K+1}$$

$$= \gamma\frac{(2\gamma^{-1}-\lfloor\gamma^{-1}\rfloor)(\lfloor\gamma^{-1}\rfloor+1)}{2(K+1)} \qquad \text{if } t > \lfloor\gamma^{-1}\rfloor.$$

Now

$$(40) \qquad \Pr(t) = \frac{\binom{K-t}{L-1}}{\binom{K}{L}}.$$

Using (38) - (40), we get the expected value of d_i.

$$\mathrm{Exp}(d_i) = \frac{1}{(K+1)\binom{K}{L}} \sum_{t=1}^{\lfloor\gamma^{-1}\rfloor} t\binom{K-t}{L-1} - \frac{\gamma}{2(K+1)\binom{K}{L}} \sum_{t=1}^{\lfloor\gamma^{-1}\rfloor} (t-1)t\binom{K-t}{L-1}$$

$$+\ \gamma\frac{(2\gamma^{-1}-\lfloor\gamma^{-1}\rfloor)(\lfloor\gamma^{-1}\rfloor+1)}{2(K+1)\binom{K}{L}}\sum_{t>\lfloor\gamma^{-1}\rfloor}\binom{K-t}{L-1}.$$

Now, as may be inductively verified,

$$\binom{L-1}{L-1}+\binom{L}{L-1}+\dots+\binom{A}{L-1}=\binom{A+1}{L}$$

$$\binom{K-1}{L-1}+2\binom{K-2}{L-1}+\dots+A\binom{K-A}{L-1}=\binom{K+1}{L+1}-A\binom{K-A}{L}-\binom{K+1-A}{L+1}$$

$$1.2\binom{K-2}{L-1}+2.3\binom{K-3}{L-1}+\dots+(A-1)A\binom{K-A}{L-1}=2\binom{K+1}{L+2}-A(A+1)\binom{K-A}{L}$$
$$-\ 2(A+1)\binom{K-A}{L+1}-2\binom{K-A}{L+2}.$$

Therefore.

$$\mathrm{Exp}(d_i)=\frac{1}{(K+1)\binom{K}{L}}\left[\binom{K+1}{L+1}-\lfloor\gamma^{-1}\rfloor\binom{K-\lfloor\gamma^{-1}\rfloor}{L}-\binom{K+1-\lfloor\gamma^{-1}\rfloor}{L+1}\right]$$
$$-\ \frac{1}{2(K+1)\binom{K}{L}}\left[2\binom{K+1}{L+2}-\lfloor\gamma^{-1}\rfloor(\lfloor\gamma^{-1}\rfloor+1)\binom{K-\lfloor\gamma^{-1}\rfloor}{L}\right.$$
$$\left.-2(\lfloor\gamma^{-1}\rfloor+1)\binom{K-\lfloor\gamma^{-1}\rfloor}{L+1}-2\binom{K-\lfloor\gamma^{-1}\rfloor}{L+2}\right]$$
$$+\ \frac{(2\gamma^{-1}-\lfloor\gamma^{-1}\rfloor)(\lfloor\gamma^{-1}\rfloor+1}{2(K+1)\binom{K}{L}}\binom{K-\lfloor\gamma^{-1}\rfloor}{L}.$$

Now $\frac{\binom{K+1}{L+1}}{(K+1)\binom{K}{L}}=\frac{1}{L+1}$, $\frac{\gamma\binom{K+1}{L+2}}{(K+1)\binom{K}{L}}=\frac{(k-L)(K-L)}{K(L+1)(L+2)}$ and $\binom{K-\lfloor\gamma^{-1}\rfloor}{L}$ $\le e^{-\lfloor\gamma^{-1}\rfloor L/K}\binom{K}{L}$. Hence, for α and β fixed, where α is close to 1 and β is large, and for $k \to \infty$,

$$\mathrm{Exp}(d_i) = \left[\frac{1}{\alpha k} - \left(\frac{1+O(\gamma)}{(1-\alpha)k} + \frac{1}{\alpha k}\right)\binom{K-\lfloor\gamma^{-1}\rfloor}{L}\right] (1+O(\tfrac{1}{k}))$$

$$- \left[\frac{(1-\alpha)(\beta-\alpha)}{\beta\alpha^2 k} - \left(\frac{1+O(\gamma)}{2(1-\alpha)k} + \frac{1+O(\gamma)}{\alpha k} + \frac{1-\alpha}{\alpha^2 k}\right)\binom{K-\lfloor\gamma^{-1}\rfloor}{L}(1+O(\tfrac{1}{k}))\right.$$

$$\left. + \frac{1+O(\gamma)}{2(1-\alpha)k}\binom{K-\lfloor\gamma^{-1}\rfloor}{L}\right](1+O(\tfrac{1}{k}))$$

$$= \frac{1}{k}\left[\frac{1}{\alpha} - \frac{(1-\alpha)(\beta-\alpha)}{\beta\alpha^2} + o\left(\frac{1}{1-\alpha} e^{-\frac{\alpha}{1-\alpha}}\right)\right] (1+O(\tfrac{1}{k}))$$

$$= \frac{1}{k}\left[1 - (\tfrac{1}{\alpha}-1)^2 + \frac{1-\alpha}{\alpha\beta} + o\left(\frac{1}{1-\alpha} e^{-\frac{\alpha}{1-\alpha}}\right)\right] (1+O(\tfrac{1}{k})).$$

Let $\beta = \frac{2\alpha}{1-\alpha}$. Then

$$\mathrm{Exp}(d_i) = \frac{1}{k}\left[1-\tfrac{1}{2}(\tfrac{1}{\alpha}-1)^2 + o\left(\frac{1}{1-\alpha} e^{-\frac{\alpha}{1-\alpha}}\right)\right] (1+O(\tfrac{1}{k})).$$

Now $(\frac{1}{\alpha} - 1)^2 / \frac{1}{1-\alpha} e^{-\frac{\alpha}{1-\alpha}} \to \infty$ as $\alpha \to 1$. Thus, by choosing α close to 1, we get

$$\mathrm{Exp}(d_i) \le \frac{1}{k}\left(1 - \tfrac{1}{3}(\tfrac{1}{\alpha} - 1)^2\right).$$

Applying Lemma 1 with $\varepsilon = \left(\frac{1}{\log n}\right)^{1/2}$, we get

$$z_{LP}(J_o, J_1) \le \frac{n}{k}\left(1 - \tfrac{1}{3}(\tfrac{1}{\alpha} - 1)^2\right)(1+o(1))$$

with probability at least $1 - e^{-\frac{nB}{3k \log n}}$ where $B = 1 - \frac{1}{3}(\frac{1}{\alpha}-1)^2$. We have to branch for all $|J_1| \le \alpha k$ and $|J_o| \le n-(\beta-\alpha)k$ with probability at least

$1 - \binom{n}{\beta k}\binom{n}{\alpha k} e^{-\frac{nB}{3k \log n}} \to 1$ since $\binom{n}{\beta k}\binom{n}{\alpha k} \le n^{(\alpha+\beta)k}$ and $\frac{n}{k^2 \log n} \to \infty$. This proves that, almost surely, (37) holds. As a consequence, the number of branches in the search tree is at least $\binom{n-\beta k}{\alpha k} = n^{(1-o(1))k}$.

□

6. Computational Experience

The previous sections provide asymptotic results as $n \to \infty$. In this section, we report our computational experience with medium-size k-median problems for the four probabilistic models introduced earlier. This computational experience is based on the solution of about 3300 random problems with $n = 50$ points and an additional 950 random problems with $n = 100$ points. The description of these problems is given later.

For each problem we computed z_{IP} and z_{LP}. The value of z_{LP} was obtained by solving a Lagrangian dual by subgradient optimization as explained in [3]. In the process of computing z_{LP}, this algorithm generates a feasible solution at each subgradient iteration. Of course, if it happens that the value of the best feasible solution generated equals z_{LP}, the algorithm terminates since, then, $z_{IP} = z_{LP}$. For most of the test problems with no gap $z_{IP} - z_{LP}$, the algorithm terminated in less than 100 subgradient iterations, due to the above stopping criterion. If, after 100 subgradient iterations, there was still a gap between the best feasible solution (an upper bound on z_{IP}) and the best Lagrangian relaxation (a lower bound on z_{LP}), we resorted to branch and bound to find z_{IP}. When the subgradient algorithm clearly converged to a value different from z_{IP}, we accepted it as showing that $z_{IP} \neq z_{LP}$. In the cases where the subgradient algorithm converged to a value close to z_{IP} we used the simplex algorithm to compute z_{LP}. This allowed us to settle cases where there was a very small but positive gap $z_{IP} - z_{LP}$.

Among the 4250 test problems that we generated we found about 3700 such that $z_{IP} = z_{LP}$ and about 550 with a gap $z_{IP} - z_{LP}$. Now we give a detailed description of these results.

The first set of experiments involves Euclidean problems. We decided to test whether approximating the Euclidean distances had an influence on the gap $z_{IP} - z_{LP}$, since we suspected that data accuracy might be partly responsible for the discrepancy between the computational experience previously reported in the literature, namely few test problems were found to have gaps ([2], [3], [6], [10], [11], [19], [20], [23], [24]), and the results of Section 2 stating that asymptatically most instances should have small but positive gaps. To our surprise, data accuracy had little influence except maybe for the possibility that a very coarse approximation produces harder k-median problems. (These problems are more combinatorial, often have alternate optimal solutions and, in our experience, optimality was harder to prove). We generated 10 problems, each with 50 points occurring at random in the unit square. Then, for i = 1,2,3,4 and 5, we multiplied each point coordinate by 10^i and rounded it to the closest integer value. The Euclidean distances were then computed and rounded to the closest integer. The k-median problem and its LP relaxation were solved for each $2 \le k \le 10$ and $1 \le i \le 5$. For each such pair i,k, Table 1 reports the number of problems (out of 10) with a gap $z_{IP} - z_{LP}$.

k / i	2	3	4	5	6	7	8	9	10	Total (out of 90)
1	0	2	0	2	2	1	0	0	0	7
2	0	1	0	0	0	2	0	0	1	4
3	0	1	0	1	0	0	2	0	0	4
4	0	1	0	0	1	0	2	0	0	4
5	0	1	0	0	1	0	2	0	0	4
Total (out of 50)	0	6	0	3	4	3	6	0	1	23 (out of 450)

Table 1 Euclidean model with n = 50.
Number of instances with a gap.

The same two problems were responsible for all the gaps. The average value of $\frac{z_{IP}-z_{LP}}{z_{IP}}$ over the instances that had a gap was approximately 1.5% for i = 1, .4% for i = 2 and .1% for i = 3,4 and 5. Overall, the fraction of instances with a gap was about 5%. This is consistent with the computational experience reported in the literature. Clearly, the asymptotic behavior described in Section 2 is not felt for problems with n = 50 points. It would be interesting to repeat the computational experiment for Euclidean k-median problems with about n=1000 points. Unfortunately our computer budget did not allow to do this.

The second set of experiments involves random trees. We generated 100 random trees, 50 of them with n = 50 nodes and the other 50 with n=100 nodes, using the method described in Even [7]. First we assumed that all edge lengths were equal to 1 in the trees, and we solved the k-median problem and the LP relaxation for $2 \leq k \leq 11$ in each tree. For each pair n,k, Table 2 reports the number of problems (out of 50) with a gap $z_{IP} - z_{LP}$.

k n	2	3	4	5	6	7	8	9	10	11	Total (out of 500)
50	0	2	2	0	7	1	1	1	2	2	18
100	1	2	1	4	0	2	2	2	2	1	17
Total (out of 100)	1	4	3	4	7	3	3	3	4	3	35 (out of 1000)

Table 2 Tree model with unit edge lengths.
Number of instances with a gap.

We also computed z_{IP} and z_{LP} for the same 100 trees assuming non-unit edge lengths. In this experiment, the nodes of the tree were random points in the unit square and the length of an edge in the tree was the Euclidean distance between its endpoints rounded using the scheme explained earlier with $i = 1$. The distance between two nodes of the tree was the length of the unique path joining them. Table 3 reports these results.

n \ k	2	3	4	5	6	7	8	9	10	Total (out of 450)
50	0	2	5	6	3	3	0	3	1	23
100	0	2	1	6	1	3	2	5	1	21
Total (out of 100)	0	4	6	12	4	6	2	8	2	44 (out of 900)

Table 3 Tree model with non-unit edge lengths.
Number of instances with a gap.

We did not find a significant difference in difficulty between the two tree models. Overall, the fraction of instances with a gap was about 4%.

Our third set of experiments involves random graphs. First we report the results when the edge lengths are equal to 1. Starting from a random tree on 50 nodes, we generated a sequence of graphs, adding 50 random edges at a time to the previous graph. Table 4 contains the value of z_{LP} and z_{IP} for each graph and $2 \leq k \leq 10$. Only one figure means that $z_{IP} = z_{LP}$. Note that when $z_{IP} = z_{LP} = n-k$ for some graph, it contains a dominating set and therefore every subsequent graph in the sequence also does.

k number of edges	2	3	4	5	6	7	8	9	10
49	139	114	98	88	80	72	65	59	54
99	89	77	68.5/69	62	57	52	48	44.5/46	42
149	77	69	62	55.5/57	50	46	43	41	40
199	72	63	55	48	45	43	42	.	.
249	72	61	52	46	44/45	.	.	.	.
299	69	56	48	46	44	.	.	.	.
349	65	52.5/54	48	45/46	.	.	.	.	.
399	62	50	47/48	45	.	.	.	.	.
449	61	49	46/47	.	.	.	.	.	.
499	58	47.5/48	46/47	.	.	.	.	.	.
549	56	48	46	.	.	.	.	.	.
599	54	47/48	.	.	.	.	.	.	.
649	52	47	.	.	.	.	.	.	.
699	51	.	.	.	.	.	.	.	.
749	50	.	.	.	.	.	.	.	.
799	48.5/49	.	.	.	.	.	.	.	.
849	48/49	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
1199	48	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

<u>Table 4</u> Graphical model with unit edge lengths.

Value of z_{LP} and z_{IP}.

Among the instances where a dominating set did not exist, about 28% had a gap.

Next we turn to the graphical model with non-unit edge lengths. We started from 10 random trees on n = 50 nodes. We then added random edges, 50 at a time, until the graphs contained 849 edges. The edge lengths were computed using the same scheme as earlier. Namely, the nodes were assigned random integer coordinates in a square of size 10x10 and the length of an edge was the Euclidean distance between its two endpoints, rounded to the closest integer. The distance between two nodes of the graph was taken to be the length of the shortest path joining them in the graph. Table 5 reports the number of instances with a gap (out of 10), as a function of the number of edges in the graph and k.

k number of edges	2	3	4	5	6	7	8	9	10	Total (out of 90)
49	0	1	1	0	0	0	0	0	1	3
99	1	1	1	2	3	1	0	1	1	11
149	2	1	2	2	1	0	0	0	0	8
199	1	2	1	0	1	0	2	1	2	10
249	1	2	2	1	1	0	1	3	1	12
299	2	1	2	2	1	2	1	1	1	13
349	2	2	4	1	5	1	0	3	2	20
399	1	3	2	0	2	1	0	1	1	11
449	3	2	2	1	1	2	2	1	0	14
499	0	1	1	2	1	1	1	1	0	8
549	1	1	4	0	0	1	2	1	1	11
599	1	1	0	2	2	2	2	0	2	12
649	2	0	1	2	0	0	3	1	1	10
699	0	2	2	1	0	2	2	0	1	10
749	0	1	1	0	1	2	1	1	1	8
799	1	1	0	1	1	1	0	2	3	10
849	0	0	2	0	2	1	0	2	3	10
Total (out of 170)	18	22	28	17	22	17	17	19	21	181 (out of 1530)

Table 5 Graphical model with non-unit edge lengths.
Number of instances with a gap.

For this model, the fraction of instances with a gap was about 12%. The average of $\frac{z_{IP}-z_{LP}}{z_{IP}}$ taken over the instances with a gap was less than 1%.

Note that the first line of Table 1 corresponds to the case of the graphical model where the number of edges is $\frac{49\text{x}50}{2} = 1250$ and, as such, could be added as a line of Table 5.

Finally, the fourth set of experiments deals with the uniform cost model. We generated 30 problems with random integer costs. In the first 10 problems the costs were in the range [1,10], in the next 10 in the range [1,100] and in the last 10 in the range [1,1000]. For each problem the values of z_{IP} and z_{LP} were computed for $2 \leq k \leq 10$. For each range and value of k, Table 6 contains the number of instances with a gap (out of 10).

k range	2	3	4	5	6	7	8	9	10	Total (out of 90)
10	10	10	10	10	10	10	10	6	4	80
100	10	10	10	10	10	10	10	8	5	83
1000	10	10	10	10	10	10	10	10	6	86
Total (out of 30)	30	30	30	30	30	30	30	24	15	249 (out of 270)

Table 6 Uniform cost model with n = 50.
Number of instances with a gap.

For this model, the fraction of instances with a gap was about 92% overall, 100% for $k \leq 8$. This fits well with the results of Section 5. The value of the ratio $\frac{z_{IP}-z_{LP}}{z_{IP}}$ was much larger than in the other models. It reached 18% for one of the problems with costs taken in the range [1,1000] and

$k = 3$. Note, however, that this is still far below the asymptotic value of 33% predicted by Theorem 8 when $k = 3$.

7. The Simple Plant Location Problem

Although we proved our probabilistic results for the k-median problem, they can also be useful for the SPLP. To define an instance of SPLP, we need fixed costs f_j, $j=1,\ldots,n$, in addition to the distances d_{ij}, $1 \le i, j \le n$. For simplicity, we assume in this section that the fixed costs f_j are all identical, say $f_j = f$.

Theorem 10 Consider the Euclidean model in the plane and assume that $n^{\varepsilon-1/2} \le f \le n^{1-\varepsilon}$ for some fixed $\varepsilon > 0$. Then, for the SPLP,

$$\frac{z_{IP}-z_{LP}}{z_{IP}} \sim .00189255\ldots \qquad \text{almost surely.}$$

Proof. In this proof, z_{IP} and z_{LP} denote the optimum values of SPLP and its linear programming relaxation respectively. The solutions of the corresponding k-median problem (with same d_{ij}'s) and its relaxation are denoted by $z_{IP}(k)$ and $z_{LP}(k)$ respectively.

By definition $z_{LP} = \min\limits_{k} (z_{LP}(k) + kf) = \min(z_1,z_2,z_3)$, where

$$z_1 = \min_{k<\omega} (z_{LP}(k) + kf),$$

$$z_2 = \min_{\omega \le k \le \frac{n}{\omega \log n}} (z_{LP}(k) + kf), \text{ and}$$

$$z_3 = \min_{k>\frac{n}{\omega \log n}} (z_{LP}(k) + kf).$$

First we compute z_2. From the proof of Theorem 2,

$$\Pr(z_{LP} \notin [\frac{2n}{3\sqrt{k\pi}}(1-o(1)), \frac{2n}{3\sqrt{k\pi}}(1+o(1))]\ ) = O(ne^{-2\omega^{1/3}\log n/9})$$

and so

$$z_2 \leq \min_{\omega \leq k \leq \frac{n}{\omega \log n}} \{\frac{2n}{3\sqrt{k\pi}}(1+o(1)) + k\,f\} \quad \text{almost surely.}$$

Let $\alpha = \frac{2}{3\sqrt{\pi}}$. The minimum of the function $\frac{\alpha n}{\sqrt{k}} + kf$ is attained when $k = (\frac{\alpha n}{2f})^{2/3}$. Note that, given our assumptions on f, this value is in the range $[\omega, \frac{n}{\omega \log n}]$ for a suitable ω, say $\omega = \log n$. The minimum value of the function is $(\frac{27}{4}\alpha^2 n^2 f)^{1/3}$. Therefore

$$z_2 = (\frac{27}{4}\alpha^2 n^2 f)^{1/3}(1+o(1)) \quad \text{almost surely.}$$

Now consider z_3. With our choice of $\omega = \log n$, we have $k > \frac{n}{(\log n)^2}$. Therefore, almost surely,

$$\begin{aligned} z_3 &\geq \frac{n}{(\log n)^2} f \\ &= \frac{n^{1/3} f^{2/3}}{(\log n)^2} \frac{z_2}{(\frac{27}{4}\alpha^2)^{1/3}}(1+o(1)) \geq z_2. \end{aligned}$$

Finally consider z_1. For all $k < \log n$, we have $z_{LP}(k) \geq z_{LP}(\log n)$. Therefore $z_1 \geq z_{LP}(\log n)$. This implies that, almost surely,

$$z_1 \geq \frac{2n}{3\sqrt{\pi}\log n}(1+o(1)) = c\,\frac{n^{1/3} f^{-1/3}}{(\log n)^{1/2}} z_2(1+o(1)) \geq z_2,$$

where c is a constant.

We have just proved that

$$z_{LP} \sim \left(\frac{27}{4}\alpha^2 n^2 f\right)^{1/3} \text{ almost surely.}$$

Similarly, $z_{IP} = \min_k (z_{IP}(k) + kf)$. Following the proof of Papadimitriou [22], we can show that

(41) $z_{IP} = \min_k \left(\frac{\beta n}{\sqrt{k}}(1+o(1)) + fk\right)$ almost surely,

where $\beta = .3771967\ldots$. The minimum in (41) is achieved when $k = \left(\frac{\beta n}{2f}\right)^{2/3}$ and its value is $\left(\frac{27}{4}\ \beta^2 n^2 f\right)^{1/3}(1+o(1))$.

So $$\frac{z_{IP}-z_{LP}}{z_{IP}} \sim \frac{\beta^{2/3}-\alpha^{2/3}}{\beta^{2/3}} \text{ almost surely.}$$

□

Similarly, the next result can be shown using the proof of Theorem 8.

<u>Theorem 11</u> Consider the uniform cost model and assume that $n^{\varepsilon-1} \le f \le n^{1-\varepsilon}$ for some fixed $\varepsilon > 0$. Then

$$\frac{z_{IP}-z_{LP}}{z_{IP}} \sim 1 - \frac{\sqrt{2}}{2} \text{ almost surely.}$$

8. Conclusion

The LP relaxation (1) - (4) has been widely used in branch and bound algorithms for the k-median problem and has been reported to provide a tight bound in practice. Our analysis shows that such good results can indeed be expected in a probabilistic sense for some problem instances, but we also identify other instances where the LP relaxation is almost surely not tight.

The probabilistic analysis is performed under four classical models in location theory, namely the Euclidean, network, tree and uniform cost models. For example, let $\omega = \omega(n) \to \infty$. When $\omega \le k \le \frac{n}{\omega \log n}$ in the Euclidean model, $z_{LP}/z_{IP} = .99716\ldots + o(1)$ almost surely, and when $\omega \le k \le \frac{n}{\omega \log n}$ in the uniform cost model, $z_{LP}/z_{IP} = .5 + o(1)$ almost surely.

Our computational experience confirms that large gaps occur frequently in the uniform cost model whereas only small gaps were observed with the other models.

Another aspect of the probabilistic analysis performed in Section 2, 3 and 5 is that, under various assumptions, branch and bound algorithms must almost surely expand a non-polynomial number of nodes to solve k-median problems to optimality.

Finally, we mention as open problems the questions of describing the asymptotic behavior of z_{LP}/z_{IP} as $n \to \infty$ when (i) $k \ge \frac{n}{\log n}$ in the Euclidean model, (ii) each edge of the graph has a random length d_{ij} (drawn uniformly in the interval [0,1], say) in the network and tree models, (iii) $\frac{n}{\log n} \le k \le \frac{n(e-1)}{e}$ in the uniform cost model.

References

[1] J.E. Beasley "An Solving Large p-Median Problems," Technical Report, DepartmManagement Science, Imperial College, London, England (Septemb·

[2] N. Christofides aBeasley "A Tree Search Algorithm for the p-Median Problem," Europeal of Operational Research 10 (1982), 196-204.

[3] G. Cornuejols, Mer and G.L. Nemhauser "Location of Bank Accounts to Optimize Flin Analytical Study of Exact and Approximate Algorithms," MarScience 23 (1977), 789-810.

[4] G. Cornuejols, Nemhauser and L.A. Wolsey "Worst-Case and Probabilistic Aif Algorithms for a Location Problem," Operations Research 28 (19858.

[5] G. Diehr "An Algor the p-Median Problem," Working Paper No. 191, Western Managemence Institute, University of California, Los Angeles (1972).

[6] D. Erlenkotteril-Based Procedure for Uncapacitated Facility Location," Opersearch 26 (1978), 992-1009.

[7] S. Even, Graphhms, Computer Science Press, Potomac, Maryland (1979).

[8] M.L. Fisher andchbaum "Probabilistic Analysis of the Planar k-Median Problem,tics of Operations Research 5 (1980), 27-34.

[9] R.D. Galvão Bounded Algorithm for the p-Median Problem," Operations Rese1980), 1112-1121.

[10] R.S. Garfinkel,ebe and M.R. Rao "An Algorithm for the m-Median Plant Location ITransportation Science 8 (1974), 217-236.

[11] M. Guignard andberg "Algorithms for Exploiting the Structure of the Simple Plaion Problem," Annals of Discrete Mathematics 1 (1977), 247-271.

[12] Hoeffding "Pr Inequalities for Sums of Bounded Random Variables," Jouthe American Statistical Association 58 (1963), 13-30.

[13] A. Kolen "Solvirg Problems and the Uncapacitated Plant Location Problem on Treeean Journal of Operational Research 12 (1983), 266-278.

[14] T.L. Magnanti . Wong "Accelerating Benders Decomposition: Algorithmic Enl and Model Selection Criteria," Operations Research 29 (1984.

[15] R.E. Marsten "An Algorithm for Finding Almost All of the Medians of a Network," Discussion Paper No. 23, The Center for Mathematical Studies in Econometrics and Management Science, Northwestern University, Evanston, Illinois (1972).

[16] L.P. Mavrides "An Indirect Method for the Generalized k-Median Problem Applied to Lock-Box Location," Management Science 25 (1979), 990-996.

[17] P.B. Mirchandani, A. Oudjit and R.T. Wong "Locational Decisions on Stochastic Multidimensional Networks" (1983).

[18] C. Mukendi "Sur l'implantation d'equipement dans un reseau: le probleme de m-centre", Thesis, University of Grenoble, France (1975).

[19] J.M.Mulvey and H.L. Crowder "Cluster Analysis: An application of Lagrangian Relaxation," Management Science 25 (1979), 329-340.

[20] S.C.Narula, U.I. Ogbu and H.M. Samuelsson "An Algorithm for the p-Median Problem," Operations Research 25 (1977), 709-713.

[21] G.L. Nemhauser and L.A. Wolsey "Maximizing Submodular Set Functions: Formulations and Analysis of Algorithms," Annals of Discrete Mathematics 11 (1981), 279-301.

[22] C.H. Papadimitriou "Worst-Case and Probabilistic Analysis of a Geometric Location Problem," SIAM Journal on Computing 10 (1981), 542-557.

[23] Ch. S. ReVelle and R.W. Swain "Control Facilities Location," Geographical Analysis 2 (1970), 30-42.

[24] L. Shrage "Implicit Representation of Variable Upper Bounds in Linear Programming," Mathematical Programming Study 4 (1975), 118-132.

[25] W. F. Stout, Almost Sure Convergence, Academic Press, New York (1974).

[26] E. Zemel "Probabilistic Analysis of Geometric Location problems," SIAM Journal on Algebraic and Discrete Methods 6, (1985), 189-200.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
MSRR 527

2. GOVT ACCESSION NO.
AD-A173 701

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
PROBABILISTIC ANALYSIS OF A RELAXATION
FOR THE k-MEDIAN PROBLEM

5. TYPE OF REPORT & PERIOD COVERED
Technical Report 6/86

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Sang Ahn
Colin Cooper
Gerard Cornuejols
Alan Frieze

8. CONTRACT OR GRANT NUMBER(s)
N00014-85-K-0198
NR 047-048

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research (Code 434)
Arlington, VA 22217

12. REPORT DATE
6/86 (Rev.)

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper provides a probabilistic analysis of the so-called "strong" linear programming relaxation of the k-median problem. The analysis is performed under four classical models in location theory, the Euclidean, network, tree and uniform cost models. For example, we show that, for the Euclidean model and $\log n \geq k \geq n/(\log n)^2$, the value of the relaxation is almost surely within .3 percent of the optimum k-median value. A similar analysis is performed for the other models. We also show that, under various assumptions, branch and bound

(over)

algorithms that use this relaxation as a bound must almost surely expand a non-polynomial number of nodes to solve the k-median problem of optimality. Finally, we report extensive computational experiments. As predicted by the probabilistic analysis, the relaxation was not as tight for the problem instances drawn from the uniform cost model as for the the other models.

END

12-86

DTIC